# PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2003-030235

(43)Date of publication of application : 31.01.2003

(51)Int.Cl. G06F 17/30

(21)Application number : 2001-212184

(22)Date of filing : 12.07.2001

(71)Applicant : CASIO COMPUT CO LTD

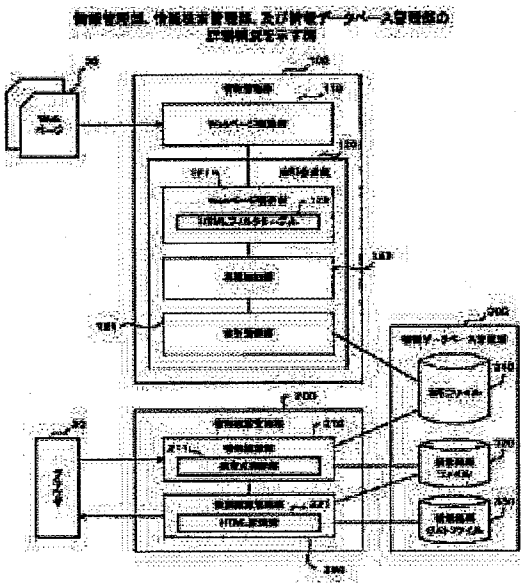
(72)Inventor : TERADA TOSHIHITO

## (54) SYSTEM AND METHOD FOR RETRIEVING INFORMATION AND PROGRAM

(57)Abstract:

PROBLEM TO BE SOLVED: To make information retrieval results provided by a retrieval engine more appropriate to a retrieval target of an information retriever.

SOLUTION: A word extracting part 123 extracts a word from a character string included in a Web page 20. An index registering part 124 acquires location information of a link destination about an extracted character string with a hyperlink embedded therein among extracted character strings, associates the location information of the Web page 20 and the location information of the link destination with the word and registers the word in an index file 310. An information retrieving part 210 retrieves the index file 310, acquires the location information associated with the word representing a retrieval object and stores the location information in a retrieval result file 320. A retrieval result managing part 220 sorts information of the retrieval result file 320, prepares an HTML file representing location information with a high priority given of link destinations to which many hyperlinks are set up and provides a browser 30 with the HTML file.



* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

---

[Claim(s)]
[Claim 1]A word contained in document information characterized by comprising the following currently released on a communication network, An index file which matches word related position information which shows a document information position in which it is the information which shows a logical position on this communication network, and information relevant to this word exists is prepared, A system which presents word related position information corresponding to a word which searches this index file based on a word showing an object of search, and expresses this retrieval object.
An extraction means to extract a word from a character string contained in said document information.
Reference destination reference-by-location speciality stage which acquires this reference destination position information about that to which an attribute which shows that it has the reference destination position information which is information which shows said position about a reference destination where information relevant to this character string is provided among said character strings is given.
A word extracted by said extraction means, A registration means to match with word related position information which consists of document position information which shows said position about document information in which a character string of extraction origin of this word was contained, and said reference destination position information about this character string to which said attribute is given, and to register with said index file.
A search means to acquire word related position information which searches said index file based on a word showing said retrieval object, and is matched with this word from this index file, and a presenting means which gives priority to and presents said reference destination position information among word related position information acquired by said search means.

[Claim 2]When said word related position information and reference destination position information which show said same position are included in word related position information

acquired by said search means, said presenting means, The information retrieval system according to claim 1 showing this word related position information as priority is given to what has many numbers acquired as reference destination position information among these word related position information.

[Claim 3]Said document information is described by Page Description Language expressing a Web page, and said reference destination reference-by-location speciality stage, The information retrieval system according to claim 1 characterized by what information which shows said position of a link destination in a hyperlink currently embedded at said character string is acquired for as said reference destination position information.

[Claim 4]Registration which matches this document position information with a character string which is the title by which said registration means is given to document information said position is indicated to be to said index file using said document position information, And registration which matches this reference destination information with a character string where a hyperlink to said link destination where said position is shown by said reference destination position information is embedded is performed, . Show this word related position information using this character string that shows a link to said position which said presenting means is the character string where a hyperlink based on matching registered into said index file was embedded, and is shown by said word related position information. The information retrieval system according to claim 3 characterized by things.

[Claim 5]Match with word related position information characterized by comprising the following, and it registers with said index file, Word related position information which searches said index file based on a word showing said retrieval object, and is matched with this word is acquired from this index file, An information retrieval method characterized by what said reference destination position information is given priority to and shown for among word related position information acquired by said search.

A word contained in document information currently released on a communication network. An index file which matches word related position information which shows a document information position in which it is the information which shows a logical position on this communication network, and information relevant to this word exists is prepared, It is the method of showing word related position information corresponding to a word which searches this index file based on a word showing an object of search, and expresses this retrieval object, Extract a word from a character string contained in said document information, and Inside of said character string, About that to which an attribute which shows that it has the reference destination position information which is information which shows said position about a reference destination where information relevant to this character string is provided is given. Document position information which shows said position about document information in which a character string of extraction origin of this word was contained in a word which acquired this reference destination position information and was extracted by said extraction, and said reference destination position information about this character string to which said attribute is

given.

[Claim 6]Processing which is matched with word related position information characterized by comprising the following, and is registered into said index file, Processing which acquires word related position information which searches said index file based on a word showing said retrieval object, and is matched with this word from this index file, A program for making processing which gives priority to and presents said reference destination position information among word related position information acquired by said search perform to a computer.
A word contained in document information currently released on a communication network by performing a computer.
An index file which matches word related position information which shows a document information position in which it is the information which shows a logical position on this communication network, and information relevant to this word exists is prepared, Processing which extracts a word from a character string which is a program for making processing which presents word related position information corresponding to a word which searches this index file based on a word showing an object of search, and expresses these conditions perform to this computer, and is contained in said document information.
Processing which acquires this reference destination position information about that to which an attribute which shows that it has the reference destination position information which is information which shows said position about a reference destination where information relevant to this character string is provided among said character strings is given.
Document position information which shows said position about document information in which a character string of extraction origin of this word was contained in a word extracted by said extraction, and said reference destination position information about this character string to which said attribute is given.

---

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

---

---

[Detailed Description of the Invention]
[0001]
[Field of the Invention]This invention relates to the art of enabling it to provide the information which agreed appropriately by the demand, to the demand of search especially about the art of retrieving information.
[0002]
[Description of the Prior Art]In recent years, the number of the Web pages provided by the WWW (WorldWideWeb) system on the Internet is continuing increasing explosively by the spread of the Internet. On the Internet, many search engines which provide the service which retrieves the information made into the purpose out of this huge information are established.
[0003]There are some which are called the robot type as one of the methods which collects the information on a network of a search engine. In a robot type search engine, the robot program called a spider or a crawler is started periodically, Automatic collection of the HTML (HyperText MarkupLanguage) file expressing the Web page currently exhibited on the Internet is performed. When information retrieval is performed and the information retrieval person using a search engine gives a closely related keyword to the target information at a search engine site, Processing which extracts that in which the keyword was contained from the collected file is performed, and an information retrieval person is provided with the list of Web pages which are the keyword and which are contained as search results with the information which shows the logical position on the Internet about the Web page.
[0004]
[Problem(s) to be Solved by the Invention]Generally, since the robot type search engine is performing [ no ] automatically processings of a to [ from collection of information / offer of search results ] by computer and operation of the information by judgment of human being intervenes, the arrangement about the quality of the genre to which the collected information belongs, or its information is not made. Therefore, if search by coincidence of a mere keyword was performed on the occasion of search of information, a Web page including important

information is buried in search results, or. Or there were not few cases where it will be mostly contained in search results only in about the Web page in which what is called a search noise, i.e., the low information on usefulness, is contained.

[0005]Making more suitable the result of the information retrieval which a search engine provides in view of the above problem to an information retrieval person's retrieval object is the issue which this invention tends to solve.

[0006]

[Means for Solving the Problem]A word contained in document information to which this invention is opened on a communication network, An index file which matches word related position information which shows a document information position in which it is the information which shows a logical position on this communication network, and information relevant to this word exists is prepared, It is premised on a system or a method of showing word related position information corresponding to a word which searches this index file based on a word showing an object of search, and expresses this retrieval object.

[0007]And in an information retrieval system which is one of the modes of this invention. An extraction means to extract a word from a character string contained in said document information, About that to which an attribute which shows that it has the reference destination position information which is information which shows said position about a reference destination where information relevant to this character string is provided among said character strings is given. A word extracted by reference destination reference-by-location speciality stage which acquires this reference destination position information, and said extraction means, A registration means to match with word related position information which consists of document position information which shows said position about document information in which a character string of extraction origin of this word was contained, and said reference destination position information about this character string to which said attribute is given, and to register with said index file, A search means to acquire word related position information which searches said index file based on a word showing said retrieval object, and is matched with this word from this index file, SUBJECT mentioned above by constituting so that it may have a presenting means which gives priority to and presents said reference destination position information among word related position information acquired by said search means is solved.

[0008]For example, in a technical paper, especially a paper currently referred to by many other papers can be considered to be what has high importance. This invention is based on this view and it is considered that position information which shows information currently referred to from other document information is more suitable compared with a thing without that right. That is, it considers that reference destination position information is more important than mere document position information in word related position information, and is made to make reference destination position information show preferentially a word related position information presenting means. Since reference destination position information which shows a

reference destination which this character string in which a word which is a character string in document information currently released on a communication network by carrying out like this, and agrees in a search condition is contained is referring to is shown more preferentially than mere document position information, A result of information retrieval will become more suitable to an information retrieval person's retrieval object.

[0009]In an information retrieval system concerning this invention mentioned above, said presenting means, When said word related position information and reference destination position information which show said same position are included in word related position information acquired by said search means, As priority is given to what has many numbers acquired as reference destination position information among these word related position information, it may be made to show this word related position information.

[0010]According to this composition, word related position information a position of information with more numbers referred to from other character strings is indicated to be comes to be given [ priority ] and shown. In an information retrieval system concerning this invention mentioned above, said document information, It is described by Page Description Language expressing a Web page, and said reference destination reference-by-location speciality stage, It may be made to acquire information which shows said position of a link destination in a hyperlink currently embedded at said character string as said reference destination position information, and the same operation and effect as an information retrieval system applied to this invention also by this composition are done so.

[0011]Registration which matches this document position information with a character string which is the title by which said registration means is given to document information said position is indicated to be to said index file using said document position information at this time, And registration which matches this reference destination information with a character string where a hyperlink to said link destination where said position is shown by said reference destination position information is embedded is performed, It may be made for said presenting means to present this word related position information using this character string that shows a link to said position which is the character string where a hyperlink based on matching registered into said index file was embedded, and is shown by said word related position information.

[0012]Since word related position information shown as search results is shown as a hyperlink embedded at a character string according to this composition, An information retrieval person who received the search results becomes possible [ arriving simply to a link destination ], and can acquire now information relevant to a word of a search condition easily.

[0013]An information retrieval method which is one of the modes of this invention, Extract a word from a character string contained in said document information, and Inside of said character string, About that to which an attribute which shows that it has the reference destination position information which is information which shows said position about a reference destination where information relevant to this character string is provided is given. A

word which acquired this reference destination position information and was extracted by said extraction, Match with word related position information which consists of document position information which shows said position about document information in which a character string of extraction origin of this word was contained, and said reference destination position information about this character string to which said attribute is given, and it registers with said index file, Word related position information which searches said index file based on a word showing said retrieval object, and is matched with this word is acquired from this index file, word related position information acquired by said search -- the same operation and effect as an information retrieval system concerning this invention mentioned above are acquired by giving priority to and showing said reference destination position information among them.
[0014]SUBJECT mentioned above by making a computer execute the program also in a program for making processing which consists of the same procedure as an information retrieval method concerning this invention mentioned above perform to a computer is solvable.

[0015]
[Embodiment of the Invention]Hereafter, an embodiment of the invention is described based on a drawing. Drawing 1 is a figure in which the information retrieval site which carries out this invention shows the entire configuration of the communication network which provides an information search service.
[0016]In drawing 1, the Internet 4 which is all a communication network is accessed, and data can be delivered [ the information retrieval site 1 the offer-of-information site 2a, 2b, 2c, 2d, and the user terminals 3a and 3b ] and received mutually. The information retrieval site 1 is a WWW server system which provides a robot search type information search service for the user terminal 3a and ab, is provided with the Research and Data Processing Department 100, the information retrieval Management Department 200, the information database Management Department 300, and the WWW server Management Department 400, and is constituted.
[0017]The Research and Data Processing Department 100 performs automatic collection of the information currently released on the Internet 4, and accumulates the collected information in the information database Management Department 300. The information retrieval Management Department 200 retrieves the information accumulated in the information database Management Department 300 according to the demand of the information retrieval sent via the Internet 4, and returns the result of the search to a requiring agency.
[0018]At the information database Management Department 300, accumulation of the information collected by the Research and Data Processing Department 100 and search of the information by the information retrieval Management Department 200 are performed. The processing by which the WWW server part 400 transmits the collected information which is sent via the Internet 4 to the Research and Data Processing Department 100, Processing which transmits the demand of the information retrieval sent via the Internet 4 to the information retrieval Management Department 200, and processing of sending out of a Web

page in which the information which shows the result of the information retrieval sent from the information retrieval Management Department 200 is expressed are performed.

[0019]The offer-of-information site 2a, 2b, and 2c and 2d are WWW server systems which exhibit Web pages 20a, 20b, 20c, and 20d on the Internet 4, respectively. Although four offer-of-information sites are shown in drawing 1, the number of the offer-of-information sites connected to the Internet 4 may be arbitrary.

[0020]The user terminals 3a and 3b, respectively The offer-of-information site 2a, 2b, 2c, And it is a computer which can perform the browsers 30a and 30b which are the software which peruses the Web page provided from 2 d or the information retrieval site 1, and is operated by the information retrieval person who requests search of the information currently released on the Internet 4 to the information retrieval site 1. Although two users are shown in drawing 1, the number of the user terminals connected to the Internet 4 may also be arbitrary.

[0021]These information retrieval sites 1, the offer-of-information site 2a, 2b, 2c and 2d, and the user terminals 3a and 3b, The computer by which all have standard composition, i.e., CPU which controls each component by executing a control program, The storage parts store used as the work area at the time of the memory and CPU of a control program which consist of a ROM, RAM, a magnetic storage device, etc., and make CPU control each component executing a control program, or a storage area of various data, It can also constitute using a computer provided with the input part from which various kinds of data corresponding to operation by a user is acquired, the outputting part which show a display etc. various kinds of data and of which a user is notified, and the I/F part which provides the interface function for connecting with a network.

[0022]Next, drawing 2 is explained. The figure shows the composition of the Research and Data Processing Department 100 with which the information retrieval site 1 in drawing 1 is equipped, the information retrieval Management Department 200, and the information database Management Department 300 still in detail. As shown in drawing 2, the Research and Data Processing Department 100 has the Web page collecting part 110 and the index production part 120, and is constituted, The information management retrieval part 200 is provided with the information retrieval section 210 and the search-results Management Department 220, and is constituted, and the information database Management Department 300 has the index file 310, the search-results file 320, and the search-results list file 330, and is constituted.

[0023]The Web page collecting part 110 collects Web pages 20 currently exhibited on the Internet 4. The index build part 120 registers into the index file 310 the index which can lengthen the position information on Web page 20 collected by the Web page collecting part 110, i.e., the position information which shows the logical position on the Internet 4 in which Web page 20 exists. The index build part 120 is provided with the Web page analyzing parts 121, the word extraction part 123, and the index registering part 124, and is constituted.

[0024]The Web page analyzing parts 121 create the HTML filter table 122 which makes the

unit of a record each HTML tag described by the text of the HTML file which analyzes Web page 20 and is expressing Web page 20. The word extraction part 123 extracts a word from the character string shown in the HTML filter table 122.

[0025]The relation between the word from which the index registering part 124 was extracted by the word extraction part 123, and the position information about Web page 20, And when the hyperlink (it only abbreviates to a "link" hereafter) is embedded in the word by Web page 20, the index data in which the relation between existence of a link and its word, and the position information on the Web page which is the link destination is shown is registered into the index file 310.

[0026]The information retrieval section 210 acquires the demand of the information retrieval sent from the user terminal by control of the browser 30 currently performed with one of the user terminals connected to the Internet 4 from the WWW server part 400, The search formula showing the conditions of the information retrieval is taken out, and it stores in the search formula storage 211. And the word (keyword) which searches the index file 300 and is shown in the search formula acquires the index data used as a title, and stores in the search-results file 320.

[0027]The search-results Management Department 220 stores in the search-results list file 330 the position information shown in the index data stored in the search-results file 320, and a number of a link of sum totals stretched to the position information, if search by the information retrieval section 210 is completed. And the HTML file expressing the Web page as which the search-results list which sorts the position information stored in the search-results list file 330 according to the number of links, and becomes if it is the sorted position information is displayed is created by the HTML preparing part 221. The created HTML file is addressed to the user terminal in which the browser 30 is performed, and is sent out to the Internet 4 by the WWW server part 400.

[0028]Next, an example is shown and explained about the details of processing of the collection of a Web page performed in the Research and Data Processing Department 100 which the information retrieval site 1 has, and generation of an index. Drawing 3 shows the example of Web page 20 which is opened to the Internet 4 and collected by the information retrieval site 1.

[0029]The Web page of a total of five sheets of HP1-1 and HP1-2, HP1-3, HP2-1, and HP2-2 is illustrated by drawing 3. The arrow shown in the figure shows the relation of the link. That is, it is shown that the link of HP1-2 is embedded at the character string which an "accommodation plan" of HP1-1 Becomes, for example.

[0030]The HTML sauce of HP1-1 is shown in drawing 4. The browser's 30 inspection of HTML shown in the figure (b) will display the screen shown in the figure (a). Drawing 5 is explained here. The figure is a flow chart which shows the contents of processing of the index production processing performed in the Research and Data Processing Department 100. By performing this processing, collection of a Web page and generation of an index are performed in the

Research and Data Processing Department 100.

[0031]First, in S101, only when it was distinguished whether the present date is the collection designated date of Web page 20 specified beforehand, this decision result is set to Yes and the present becomes that designated date, processing progresses to S102. Although the method of specification of this date is arbitrary, specification called the monthly final day of month end, etc. is performed, for example.

[0032]In S102, processing of a round and collection of Web page 20 currently exhibited on the Internet 4 by the Web page collecting part 110 is performed. The technique of this round and collection should just use as it is what is performed from the former by the well-known robot type search engine.

[0033]In S103, the tag form of the HTML sauce of collected Web page 20 is analyzed by the Web page analyzing parts 121, and a HTML filter table is generated by the Web page analyzing parts 121 in S104 continuing. The HTML filter table generated from HP1-1 shown in drawing 3 is shown in drawing 6. In the Web page analyzing parts 121, the HTML filter table which the HTML sauce about HP1-1 shown in drawing 4 (b) is analyzed, and is shown in drawing 6 is generated.

[0034]When the contents of processing of S103 are explained further, referring to drawing 4 (b), in the Web page analyzing parts 121. In the <BR> tag (line feed tag), it is considered that the text of the HTML sauce of an analytical object, i.e., all the character strings inserted between the start tag of <BODY> and the end tag, is a pause of a character string, and they are extracted.

[0035]In processing of Scontinuing 104, about the display which shows the selected character string and why the character string was chosen, and the thing where the link to other Web pages is embedded, the position information on the link destination is summarized as one record, and the HTML filter table 122 is generated.

[0036]If signs that the HTML filter table shown in drawing 6 from the HTML sauce shown in drawing 4 (b) is created are explained, First, the portion pinched between the start tag of the <BODY> tag and end tag which are the description parts of the text in HTML sauce, That is, the character string contained in the portion put between the <BODY> tag and the </BODY> tag is divided into four character strings which a "traffic & map" [ "it is Welcome! to the Hakone hotel", an "accommodation plan", "circumference sightseeing guidance", and ] Consist of <BR> tags.

[0037]And the classification "STRING" which shows that it is a character string where the link is not embedded is given to "it is Welcome! to the Hakone hotel" among these character strings, and one record of a HTML filter table is generated. Since the link to other Web pages is embedded, each of each character strings of an "accommodation plan", "circumference sightseeing guidance", and a "traffic & map", Classification that it is the character string where "LINK, i.e., a link," is embedded is given to these character strings, The record of the HTML filter table which consists of the character string and classification, and URL (Uniform

Resource Locator) of the link destination which is the position information on the link destination of each character string is generated for every character string of the.
[0038]In S106 which the record of the HTML filter table 122 is specified one [ at a time ] in order in the word extraction part 123, and continues in S105, It is distinguished by the word extraction part 123 whether the data in which the classification of the character string shown in the specified record is shown is either "STRING" or "LINK." And if the result of this distinction becomes in Yes, logging of the word which constitutes the character string shown in that record in S107 will be performed by the word extraction part 123. And in S108 continuing, the started word is made into a title, and the index which matched with the word of the title the title which is the page in which the word was contained, and position information is generated by the index registering part 124, and is registered into the index file 310.
[0039]On the other hand, if the result of the discrimination processing of S105 becomes in No, processing will progress to S109. In S109, it is distinguished by the word extraction part 123 whether the specification of S105 mentioned above about all the records of the HTML filter table 122 was made, and if the result of this distinction becomes in Yes, processing will progress to S110. On the other hand, if the result of this discrimination processing becomes in No, the processing which processing returned and mentioned above to S104 will be repeated.
[0040]In S111 which the record of the HTML filter table 122 is anew specified one [ at a time ] in order, and continues by the word extraction part 123 S110, It is distinguished by the word extraction part 123 whether the data in which the classification of the character string shown in the specified record is shown is "LINK." And if the result of this distinction becomes in Yes, logging of the word which constitutes the character string shown in that record in S112 will be performed by the word extraction part 123. And the data which made the group position information on the Web page of the link destination of the character string which is started word's logging origin, and its character string in S113 continuing, The word is registered by the index registering part 124 to the record in the index file 310 which is an entry, and in S114 continuing, the link flag about the data combines and the index registering part 124 registers with the record.
[0041]On the other hand, if the result of the discrimination processing of S111 becomes in No, processing will progress to S115. In S115, it is distinguished whether the specification of S110 mentioned above about all the records of the HTML filter table 122 was made, and if the result of this distinction becomes in Yes, this index production processing will be completed. On the other hand, if the result of this discrimination processing becomes in No, the processing which processing returned and mentioned above to S110 will be repeated.
[0042]Processing to the above is index production processing. Next, the processing performed by applying to S115 from S105 is further explained using the example of drawing 3. Drawing 7 shows the data structure of the index file 310 generated by the information database Management Department 300 by index production processing which was mentioned above in the case of the example of drawing 3. In drawing 7, since it will become complicated if URL is

shown as position information instead, the name of the HP1-1 grade given to each Web page shown in drawing 3 is shown.

[0043]In the following explanation, the HTML filter file about HP1-1 shown in drawing 6 by processing to S104 mentioned above shall be generated. In drawing 6, first, since the classification about the character string "it is Welcome! to the Hakone hotel" of this record is "STRING" when a top record is specified by processing of S105, the discriminated result of S106 serves as Yes, and processing progresses to S107.

[0044]In S107, logging of a word is performed from a character string "it is Welcome! to the Hakone hotel." May adopt a well-known method as processing of logging of a word, for example, what is called a morphological analysis is used, The method made into the word which acquired the canonical form of the word for the part of speech and conjugated form of the word which were started using various kinds of dictionaries, and started the word of the canonical form from the character string, What is called an N gram method that starts the word of length N mechanically in order may be adopted shifting logging of a character string of one character at a time from the head of the character string.

[0045]Here, "Hakone" and a "hotel" should be started as a word from the character string "it is Welcome! to the Hakone hotel." At the title of the Web page which made the entry each of the word "Hakone" started by processing of the front step, and the "hotel" in S108 and from which the word was extracted, i.e., here, "The Hakone hotel", The index which made "H.P.1-1" the group is generated and it registers with the index file 310 at the position information on this Web page, i.e., here. By this processing of S108, each data of each data of "Hakone" of the 1st line of the index file shown in drawing 7, "H.P.1-1", and the "Hakone hotel" and the "hotel" of the 2nd line, "H.P.1-1", and the "Hakone hotel" is registered.

[0046]Next, although the result of the discrimination processing of S109 serves as No and the record of the 2nd line of a HTML filter file is specified by processing of S105, since the classification of this record is "BR", the result of the discrimination processing of S106 serves as No. Then, the result of the discrimination processing of S109 serves as No, and the record of the 3rd line of a HTML filter file is specified by processing of S106. Since the classification of the character string "accommodation plan" of this record is "LINK", the result of the discrimination processing of S106 serves as Yes, and processing progresses to S107.

[0047]In S107, logging of a character string "accommodation plan" to a character string is performed, and a word "stay" and a "plan" are started. At the title of the Web page which made the entry these word "stay" and "plans" of each and from which that word was extracted in S108, i.e., here, "The Hakone hotel", The index which made "H.P.1-1" the group is generated and it registers with the index file 310 at the position information on this Web page, i.e., here. By this processing of S108, each data of each data of the 3rd-line "stay" of the index file shown in drawing 7, "H.P.1-1", and the "Hakone hotel" and the "plan" of the 4th line, "H.P.1-1", and the "Hakone hotel" is registered.

[0048]Processing with the same said of a character string "circumference sightseeing

guidance" and a "traffic & map" is performed hereafter, If each data from the 1st row about the entry "circumference" applied to the 9th line from the 5th line of the index file shown in drawing 7, "sightseeing", "guidance", "traffic", and a "map" to the 3rd row is registered, the result of the discrimination processing of S109 will serve as No, and processing will progress to S110.

[0049]Next, although the record of the 1st line of a HTML filter file is anew specified by processing of S110, since the classification of this record is "STRING", the result of the discrimination processing of S111 serves as No, and processing progresses to S115. Since the result of the discrimination processing of S115 serves as No, processing returns to S105, and the record of the 2nd line of a HTML filter file is specified by this processing of S105 here, but since the classification of this record is "BR", the result of the discrimination processing of S106 serves as No again.

[0050]Then, the result of the discrimination processing of S115 serves as No, and the record of the 3rd line of a HTML filter file is specified by processing of S110. Since the classification of the character string "accommodation plan" of this record is "LINK", the result of the discrimination processing of S111 serves as Yes, and processing progresses to S112.

[0051]In S112, logging of a character string "accommodation plan" to a character string is performed, and a word "stay" and a "plan" are started. The data which made the group position information "H.P.1-2" on the Web page of the link destination of the character string "accommodation plan" which is word "stay" and "plan's" which were started logging origin in S113, and its character string, The word is registered by the index registering part 124 to the record in the index file 310 which is an entry, and in S114 continuing, the link flag about the data combines and the index registering part 124 registers with the record. The 3rd-line "stay" of the index file shown in drawing 7 by this processing of S113, Each data of each data of "H.P.1-2" and an "accommodation plan" and the "plan" of the 4th line, "H.P.1-2", and an "accommodation plan" is registered, and "1" is registered into the link flag of the 7th row of each of the 3rd line and the 4th line of an index file by processing of Scontinuing 114.

[0052]Processing with the same said of a character string "circumference sightseeing guidance" and a "traffic & map" is performed hereafter, If each data from the 5th row about the entry "circumference" applied to the 9th line from the 5th line of the index file shown in drawing 7, "sightseeing", "guidance", "traffic", and a "map" to the 7th row is registered, The result of the discrimination processing of S115 serves as Yes, and the index production processing about HP1-1 is completed.

[0053]The index file shown in drawing 7 is a thing when index production processing about HP2-1 is performed following the index production processing about HP1-1. When the index record already registered when registering an index into an index file further as shown in the figure is investigated first and the same entry exists, it is made to perform additional registration to the index record about the already registered same entry, without newly generating a record.

[0054]Next, an example is shown and explained about the details of processing of the

collection of a Web page performed at the information retrieval Management Department 200 which the information retrieval site 1 has, and generation of an index. Drawing 8 is a flow chart which shows the contents of processing of the retrieval processing performed by the information management retrieval part 200. First, the contents of processing of retrieval processing are explained along with the figures.

[0055]In S201, only when it is distinguished by the information retrieval section 210 whether the search formula in which the word which are the demand of information retrieval and the conditions of that search is shown has been sent from the browser 30 and this discriminated result is set to Yes, processing progresses to S202. In the information retrieval section 210, if a search formula is sent, the search formula is stored in the search formula storage 211.

[0056]In S202, the sent search formula is analyzed by the information retrieval section 210, and the word which is a search condition is started from the search formula. Search of the entry of the index file 310 which used the search word as the key is performed by the information retrieval section 210 in the turn that the search word was started S203.

[0057]In S204, as a result of search at a front step, it is distinguished by the information retrieval section 210 whether the entry which is in agreement with a search word was discovered, if the result of this distinction becomes in Yes, processing will progress to S205, and if No becomes, processing will progress to S207. In S206 which all of position information, a character string, and a link flag are acquired from the retrieval record in which the entry which is in agreement with a search word was contained by the information retrieval section 210, and continues by it in S205, The record which consists of this acquired entry, and position information, a character string and a link flag is stored in the search-results file 320 by the information retrieval section 210.

[0058]It is distinguished [ which was mentioned above about all the search words started in S207 by the processing of S202 mentioned above ] by the information retrieval section 210 whether processing of search of S203 was performed, If the result of this distinction becomes in Yes, processing will progress to S208, if the result of this distinction becomes on the other hand in No, processing will return to S203 and processing mentioned above about the search word in which processing of search is not yet performed will be performed.

[0059]The position information applicable to all the search words started by the processing of S202 mentioned above among the position information stored in the search-results file 320 in S208 here, That is, the position information included common to all the records of the search-results file 320 is extracted from the search-results file 320 by the search-results Management Department 220.

[0060]In S209. [ whether processing at a front step was able to extract position information, and ] That is, it is distinguished by the search-results Management Department 220 whether the position information included common to all the records of the search-results file 320 existed, if the result of this distinction becomes in Yes, processing will progress to S210, and if No becomes, processing will progress to S213.

[0061]In S210, the position information extracted by the processing of S208 mentioned above and the character string which is matched and is stored in the position information in the search-results file 320 are stored in the search-results list file 330 by the search-results Management Department 220.

[0062]The number of the link flag which is matched and is stored in the position information extracted by the processing of S208 mentioned above in the search-results file 320 in S211 is calculated by the search-results Management Department 220 for every position information of the, A counting result is stored in the search-results list file 330.

[0063]In S212, the search-results list file 330 is sorted by the search-results Management Department 220 so that it may become descending of the enumerated data of the link flag calculated by the front step. In S214 which the HTML file which expresses the contents of the search-results list after sorting by a Web page in S213 based on the search-results list file 330 is created by the HTML preparing part 221, and follows, The search-results Management Department 220 sends out the HTML file addressed and created to the browser 30 which is the transmitting origin of the search formula mentioned above in the WWW server part 400, and this retrieval processing ends it.

[0064]Processing to the above is retrieval processing. Next, the case where what was shown in drawing 7 as the index file 310 is stored in the information database Management Department 300 is made into an example, and this retrieval processing is explained. First, if a search formula is sent from the browser 30, the result of distinction of S201 will serve as Yes, and logging of a search word will be performed in S202 continuing. Here, each word of "Hakone", "sightseeing", and "guidance" should be started as a search word as a result of processing of S202.

[0065]If logging of a search word is completed, processing will progress to S203, first, search of a search word "Hakone" is performed about the entry of the index file 310, and the record about the entry "Hakone" in drawing 7 is discovered. Therefore, the result of the discrimination processing of S204 serves as Yes, and processing follows it to S205.

[0066]In S205, all of position information, a character string, and a link flag are acquired from the discovered record, and the record which consists of position information, a character string, and a link flag in S206 continuing is stored in the search-results file 320. Then, although discrimination processing in S207 is performed, since processing of search of S203 is not yet performed about "guidance" among the search words "sightseeing" started by the processing of S202 mentioned above, the result of the discrimination processing of S207 serves as No, and processing returns to S203.

[0067]Henceforth, the same processing as the search word "Hakone" mentioned above about a search word "sightseeing" and "guidance" is performed, The record which an entry "sightseeing" and the record about "guidance" are discovered, and consists of the position information, character string, and the link flag and the search word "Hakone" in the record is stored in the search-results file 320 from the index file 310 shown in drawing 7.

[0068]The contents of the search-results file 320 generated by processing to the above are shown in drawing 9. After the search-results file 320 shown in this drawing 9 is generated, the result of the discrimination processing of S207 serves as Yes, and processing progresses to S208.

[0069]In S208, three, HP1-1, HP2-1, and HP2-2, are extracted as position information which extraction of the position information included common to all the records of the search-results file 320 is performed, and is included common to all the records of "Hakone", "sightseeing", and "guidance" as a result. Therefore, the result of the discrimination processing of Scontinuing 209 serves as Yes, and processing progresses to S210.

[0070]The character string which is matched and is stored in the position information in S210 in position information HP1-1 extracted, HP2-1 and HP2-2, and the search-results file 320 is stored in the search-results list file 330, In S211 continuing, the number of the link flag which is matched and is stored in each of position information HP1-1 extracted, HP2-1, and HP2-2 is calculated, respectively, and the counting result is stored in the search-results list file 330.

[0071]Drawing 10 is explained here. The figure shows the contents of the search-results list file 330, and what is shown in the figure (a) is created as the search-results list file 330 by the processing to S211 mentioned above. Since the link flag about position information HP1-1 and HP2-1 is not stored in the search-results file 320 shown in drawing 9 at all, The link flag number is set to "0" about position information HP1-1 and HP2-1 in the search-results list file 330 shown in drawing 10 (a). On the other hand, since a total of three link flags are stored about position information HP2-2 in the search-results file 320 shown in drawing 9, the link flag number about position information HP2-2 in the search-results list file 330 shown in drawing 10 (a) is set to "3."

[0072]Two character strings, "circumference sightseeing guidance" and the "Hakone peripheral guide", are stored as a character string about position information HP2-2 in the search-results list file 330 shown in drawing 10 (a). Thus, when the character strings stored in the same position information in the search-results file 320 shown in drawing 9 differ, all the different character strings are stored in the search-results list file 330.

[0073]In S212 continuing, creation of the search-results list file 330 which shows drawing 10 (a) the contents by processing to S211 mentioned above will perform sorting of the search-results list file 330 so that it may become descending of the enumerated data of a link flag. The result to which sorting based on the enumerated data of a link flag was carried out to the search-results list file 330 of drawing 10 (a) is shown in drawing 10 (b).

[0074]Then, in S213, the HTML file which the HTML file which expresses the contents of the search-results list file 330 to which sorting was performed like drawing 10 (b) by a Web page was created, and was created in S214 continuing is sent out, and this retrieval processing is completed.

[0075]The example of a screen of the Web page which shows the result of information retrieval displayed when created HTML is perused by the browser 30 is shown in drawing 11. In the

screen shown in drawing 11, to the character string of "circumference sightseeing guidance" and the "Hakone peripheral guide." The link of HP2-2 by which position information is matched with those character strings in the search-results list file 330 shown in drawing 10 (b) is embedded, Similarly, the link of HP1-1 is embedded at a character string "Hakone hotel", and the link of HP2-1 is embedded at the character string "Hakone tourist agency", respectively. Thus, in creation processing of an HTML file [ in / in the HTML preparing part 221 / S213 ], The link to the Web page by which position information is matched with each character string by the search-results list file 330 creates the HTML file currently embedded at the display of those character strings.

[0076]The control program for making it carry out to the computer which has the standard composition which mentioned above the index production processing and retrieval processing which the information site 1 was performing in the embodiment of explained this invention by the above, and the same processing is created, By making the control program read into the computer, and performing it, this invention can be carried out by such computer.

[0077]It is also possible to carry out this invention by computer by making such a control program record on the recording medium which can be read by computer, making the program read from a recording medium to a computer, and performing it. The example of the possible recording medium of reading the control program made to record by computer is shown in drawing 12. The memory storage 502, such as ROM with which the computer 501 is equipped as built-in or external attachment as a recording medium, for example as shown in the figure, and a hard disk drive, or a flexible disk, Portable good signifier recording-media 503 grades, such as MO (magneto-optical disc), CD-ROM, and DVD-ROM, can be used. A recording medium may be the memory storage 506 which is connected with the computer 501 via the network 504 and with which the computer which functions as the program server 505 is provided. In this case, the transmission signal acquired by modulating a subcarrier with the data signal expressing a control program, The control program concerned can be executed now by making it transmit through the network 504 which is a transmission medium from the program server 5055, restoring to the transmission signal received by computer 501, and reproducing a control program.

[0078]

[Effect of the Invention]According to this invention, the word which constitutes the character string contained in the document information currently opened to the index file on the communication network, By matching and registering with the word related position information which consists of document position information which shows the position of the document information in which the character string was contained, and reference destination position information which shows the position about the reference destination where the information relevant to the character string is provided. Since reference destination position information can be given priority to and shown among the word related position information acquired by the search when the index file is searched based on the word showing a retrieval object, the

result of more suitable information retrieval can be provided to an information retrieval person's retrieval object.

---

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

---

[Field of the Invention]This invention relates to the art of enabling it to provide the information which agreed appropriately by the demand, to the demand of search especially about the art of retrieving information.

---

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

---

[Description of the Prior Art]In recent years, the number of the Web pages provided by the WWW (WorldWideWeb) system on the Internet is continuing increasing explosively by the spread of the Internet. On the Internet, many search engines which provide the service which retrieves the information made into the purpose out of this huge information are established.
[0003]There are some which are called the robot type as one of the methods which collects the information on a network of a search engine. In a robot type search engine, the robot program called a spider or a crawler is started periodically, Automatic collection of the HTML (HyperText MarkupLanguage) file expressing the Web page currently exhibited on the Internet is performed. When information retrieval is performed and the information retrieval person using a search engine gives a closely related keyword to the target information at a search engine site, Processing which extracts that in which the keyword was contained from the collected file is performed, and an information retrieval person is provided with the list of Web pages which are the keyword and which are contained as search results with the information which shows the logical position on the Internet about the Web page.

---

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

---

[Effect of the Invention]The word which constitutes the character string contained in the document information currently opened to the index file on the communication network from this invention, By matching and registering with the word related position information which consists of document position information which shows the position of the document information in which the character string was contained, and reference destination position information which shows the position about the reference destination where the information relevant to the character string is provided. When the index file is searched based on the word showing a retrieval object, reference destination position information can be given priority to and shown among the word related position information acquired by the search.
Therefore, the result of more suitable information retrieval can be provided to an information retrieval person's retrieval object.

---

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

---

[Problem(s) to be Solved by the Invention]Generally, since the robot type search engine is performing [ no ] automatically processings of a to [ from collection of information / offer of search results ] by computer and operation of the information by judgment of human being intervenes, the arrangement about the quality of the genre to which the collected information belongs, or its information is not made. Therefore, if search by coincidence of a mere keyword was performed on the occasion of search of information, a Web page including important information is buried in search results, or. Or there were not few cases where it will be mostly contained in search results only in about the Web page in which what is called a search noise, i.e., the low information on usefulness, is contained.
[0005]Making more suitable the result of the information retrieval which a search engine provides in view of the above problem to an information retrieval person's retrieval object is the issue which this invention tends to solve.

---

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

---

---

[Means for Solving the Problem]A word contained in document information to which this invention is opened on a communication network, An index file which matches word related position information which shows a document information position in which it is the information which shows a logical position on this communication network, and information relevant to this word exists is prepared, It is premised on a system or a method of showing word related position information corresponding to a word which searches this index file based on a word showing an object of search, and expresses this retrieval object.
[0007]And in an information retrieval system which is one of the modes of this invention. An extraction means to extract a word from a character string contained in said document information, About that to which an attribute which shows that it has the reference destination position information which is information which shows said position about a reference destination where information relevant to this character string is provided among said character strings is given. A word extracted by reference destination reference-by-location speciality stage which acquires this reference destination position information, and said extraction means, A registration means to match with word related position information which consists of document position information which shows said position about document information in which a character string of extraction origin of this word was contained, and said reference destination position information about this character string to which said attribute is given, and to register with said index file, A search means to acquire word related position information which searches said index file based on a word showing said retrieval object, and is matched with this word from this index file, SUBJECT mentioned above by constituting so that it may have a presenting means which gives priority to and presents said reference destination position information among word related position information acquired by said search means is solved.
[0008]For example, in a technical paper, especially a paper currently referred to by many other papers can be considered to be what has high importance. This invention is based on this view and it is considered that position information which shows information currently referred to from

other document information is more suitable compared with a thing without that right. That is, it considers that reference destination position information is more important than mere document position information in word related position information, and is made to make reference destination position information show preferentially a word related position information presenting means. Since reference destination position information which shows a reference destination which this character string in which a word which is a character string in document information currently released on a communication network by carrying out like this, and agrees in a search condition is contained is referring to is shown more preferentially than mere document position information, A result of information retrieval will become more suitable to an information retrieval person's retrieval object.

[0009]In an information retrieval system concerning this invention mentioned above, said presenting means, When said word related position information and reference destination position information which show said same position are included in word related position information acquired by said search means, As priority is given to what has many numbers acquired as reference destination position information among these word related position information, it may be made to show this word related position information.

[0010]According to this composition, word related position information a position of information with more numbers referred to from other character strings is indicated to be comes to be given [ priority ] and shown. In an information retrieval system concerning this invention mentioned above, said document information, It is described by Page Description Language expressing a Web page, and said reference destination reference-by-location speciality stage, It may be made to acquire information which shows said position of a link destination in a hyperlink currently embedded at said character string as said reference destination position information, and the same operation and effect as an information retrieval system applied to this invention also by this composition are done so.

[0011]Registration which matches this document position information with a character string which is the title by which said registration means is given to document information said position is indicated to be to said index file using said document position information at this time, And registration which matches this reference destination information with a character string where a hyperlink to said link destination where said position is shown by said reference destination position information is embedded is performed, It may be made for said presenting means to present this word related position information using this character string that shows a link to said position which is the character string where a hyperlink based on matching registered into said index file was embedded, and is shown by said word related position information.

[0012]Since word related position information shown as search results is shown as a hyperlink embedded at a character string according to this composition, An information retrieval person who received the search results becomes possible [ arriving simply to a link destination ], and can acquire now information relevant to a word of a search condition easily.

[0013]An information retrieval method which is one of the modes of this invention, Extract a word from a character string contained in said document information, and Inside of said character string, About that to which an attribute which shows that it has the reference destination position information which is information which shows said position about a reference destination where information relevant to this character string is provided is given. A word which acquired this reference destination position information and was extracted by said extraction, Match with word related position information which consists of document position information which shows said position about document information in which a character string of extraction origin of this word was contained, and said reference destination position information about this character string to which said attribute is given, and it registers with said index file, Word related position information which searches said index file based on a word showing said retrieval object, and is matched with this word is acquired from this index file, word related position information acquired by said search -- the same operation and effect as an information retrieval system concerning this invention mentioned above are acquired by giving priority to and showing said reference destination position information among them.
[0014]SUBJECT mentioned above by making a computer execute the program also in a program for making processing which consists of the same procedure as an information retrieval method concerning this invention mentioned above perform to a computer is solvable.

[0015]
[Embodiment of the Invention]Hereafter, an embodiment of the invention is described based on a drawing. Drawing 1 is a figure in which the information retrieval site which carries out this invention shows the entire configuration of the communication network which provides an information search service.
[0016]In drawing 1, the Internet 4 which is all a communication network is accessed, and data can be delivered [ the information retrieval site 1 the offer-of-information site 2a, 2b, 2c, 2d, and the user terminals 3a and 3b ] and received mutually. The information retrieval site 1 is a WWW server system which provides a robot search type information search service for the user terminal 3a and ab, is provided with the Research and Data Processing Department 100, the information retrieval Management Department 200, the information database Management Department 300, and the WWW server Management Department 400, and is constituted.
[0017]The Research and Data Processing Department 100 performs automatic collection of the information currently released on the Internet 4, and accumulates the collected information in the information database Management Department 300. The information retrieval Management Department 200 retrieves the information accumulated in the information database Management Department 300 according to the demand of the information retrieval sent via the Internet 4, and returns the result of the search to a requiring agency.
[0018]At the information database Management Department 300, accumulation of the information collected by the Research and Data Processing Department 100 and search of the

information by the information retrieval Management Department 200 are performed. The processing by which the WWW server part 400 transmits the collected information which is sent via the Internet 4 to the Research and Data Processing Department 100, Processing which transmits the demand of the information retrieval sent via the Internet 4 to the information retrieval Management Department 200, and processing of sending out of a Web page in which the information which shows the result of the information retrieval sent from the information retrieval Management Department 200 is expressed are performed.

[0019]The offer-of-information site 2a, 2b, and 2c and 2d are WWW server systems which exhibit Web pages 20a, 20b, 20c, and 20d on the Internet 4, respectively. Although four offer-of-information sites are shown in drawing 1, the number of the offer-of-information sites connected to the Internet 4 may be arbitrary.

[0020]The user terminals 3a and 3b, respectively The offer-of-information site 2a, 2b, 2c, And it is a computer which can perform the browsers 30a and 30b which are the software which peruses the Web page provided from 2 d or the information retrieval site 1, and is operated by the information retrieval person who requests search of the information currently released on the Internet 4 to the information retrieval site 1. Although two users are shown in drawing 1, the number of the user terminals connected to the Internet 4 may also be arbitrary.

[0021]These information retrieval sites 1, the offer-of-information site 2a, 2b, 2c and 2d, and the user terminals 3a and 3b, The computer by which all have standard composition, i.e., CPU which controls each component by executing a control program, The storage parts store used as the work area at the time of the memory and CPU of a control program which consist of a ROM, RAM, a magnetic storage device, etc., and make CPU control each component executing a control program, or a storage area of various data, It can also constitute using a computer provided with the input part from which various kinds of data corresponding to operation by a user is acquired, the outputting part which show a display etc. various kinds of data and of which a user is notified, and the I/F part which provides the interface function for connecting with a network.

[0022]Next, drawing 2 is explained. The figure shows the composition of the Research and Data Processing Department 100 with which the information retrieval site 1 in drawing 1 is equipped, the information retrieval Management Department 200, and the information database Management Department 300 still in detail. As shown in drawing 2, the Research and Data Processing Department 100 has the Web page collecting part 110 and the index production part 120, and is constituted, The information management retrieval part 200 is provided with the information retrieval section 210 and the search-results Management Department 220, and is constituted, and the information database Management Department 300 has the index file 310, the search-results file 320, and the search-results list file 330, and is constituted.

[0023]The Web page collecting part 110 collects Web pages 20 currently exhibited on the Internet 4. The index build part 120 registers into the index file 310 the index which can

lengthen the position information on Web page 20 collected by the Web page collecting part 110, i.e., the position information which shows the logical position on the Internet 4 in which Web page 20 exists. The index build part 120 is provided with the Web page analyzing parts 121, the word extraction part 123, and the index registering part 124, and is constituted.

[0024]The Web page analyzing parts 121 create the HTML filter table 122 which makes the unit of a record each HTML tag described by the text of the HTML file which analyzes Web page 20 and is expressing Web page 20. The word extraction part 123 extracts a word from the character string shown in the HTML filter table 122.

[0025]The relation between the word from which the index registering part 124 was extracted by the word extraction part 123, and the position information about Web page 20, And when the hyperlink (it only abbreviates to a "link" hereafter) is embedded in the word by Web page 20, the index data in which the relation between existence of a link and its word, and the position information on the Web page which is the link destination is shown is registered into the index file 310.

[0026]The information retrieval section 210 acquires the demand of the information retrieval sent from the user terminal by control of the browser 30 currently performed with one of the user terminals connected to the Internet 4 from the WWW server part 400, The search formula showing the conditions of the information retrieval is taken out, and it stores in the search formula storage 211. And the word (keyword) which searches the index file 300 and is shown in the search formula acquires the index data used as a title, and stores in the search-results file 320.

[0027]The search-results Management Department 220 stores in the search-results list file 330 the position information shown in the index data stored in the search-results file 320, and a number of a link of sum totals stretched to the position information, if search by the information retrieval section 210 is completed. And the HTML file expressing the Web page as which the search-results list which sorts the position information stored in the search-results list file 330 according to the number of links, and becomes if it is the sorted position information is displayed is created by the HTML preparing part 221. The created HTML file is addressed to the user terminal in which the browser 30 is performed, and is sent out to the Internet 4 by the WWW server part 400.

[0028]Next, an example is shown and explained about the details of processing of the collection of a Web page performed in the Research and Data Processing Department 100 which the information retrieval site 1 has, and generation of an index. Drawing 3 shows the example of Web page 20 which is opened to the Internet 4 and collected by the information retrieval site 1.

[0029]The Web page of a total of five sheets of HP1-1 and HP1-2, HP1-3, HP2-1, and HP2-2 is illustrated by drawing 3. The arrow shown in the figure shows the relation of the link. That is, it is shown that the link of HP1-2 is embedded at the character string which an "accommodation plan" of HP1-1 Becomes, for example.

[0030]The HTML sauce of HP1-1 is shown in drawing 4. The browser's 30 inspection of HTML shown in the figure (b) will display the screen shown in the figure (a). Drawing 5 is explained here. The figure is a flow chart which shows the contents of processing of the index production processing performed in the Research and Data Processing Department 100. By performing this processing, collection of a Web page and generation of an index are performed in the Research and Data Processing Department 100.

[0031]First, in S101, only when it was distinguished whether the present date is the collection designated date of Web page 20 specified beforehand, this decision result is set to Yes and the present becomes that designated date, processing progresses to S102. Although the method of specification of this date is arbitrary, specification called the monthly final day of month end, etc. is performed, for example.

[0032]In S102, processing of a round and collection of Web page 20 currently exhibited on the Internet 4 by the Web page collecting part 110 is performed. The technique of this round and collection should just use as it is what is performed from the former by the well-known robot type search engine.

[0033]In S103, the tag form of the HTML sauce of collected Web page 20 is analyzed by the Web page analyzing parts 121, and a HTML filter table is generated by the Web page analyzing parts 121 in S104 continuing. The HTML filter table generated from HP1-1 shown in drawing 3 is shown in drawing 6. In the Web page analyzing parts 121, the HTML filter table which the HTML sauce about HP1-1 shown in drawing 4 (b) is analyzed, and is shown in drawing 6 is generated.

[0034]When the contents of processing of S103 are explained further, referring to drawing 4 (b), in the Web page analyzing parts 121. In the <BR> tag (line feed tag), it is considered that the text of the HTML sauce of an analytical object, i.e., all the character strings inserted between the start tag of <BODY> and the end tag, is a pause of a character string, and they are extracted.

[0035]In processing of Scontinuing 104, about the display which shows the selected character string and why the character string was chosen, and the thing where the link to other Web pages is embedded, the position information on the link destination is summarized as one record, and the HTML filter table 122 is generated.

[0036]If signs that the HTML filter table shown in drawing 6 from the HTML sauce shown in drawing 4 (b) is created are explained, First, the portion pinched between the start tag of the <BODY> tag and end tag which are the description parts of the text in HTML sauce, That is, the character string contained in the portion put between the <BODY> tag and the </BODY> tag is divided into four character strings which a "traffic & map" [ "it is Welcome! to the Hakone hotel", an "accommodation plan", "circumference sightseeing guidance", and ] Consist of <BR> tags.

[0037]And the classification "STRING" which shows that it is a character string where the link is not embedded is given to "it is Welcome! to the Hakone hotel" among these character

strings, and one record of a HTML filter table is generated. Since the link to other Web pages is embedded, each of each character strings of an "accommodation plan", "circumference sightseeing guidance", and a "traffic & map", Classification that it is the character string where "LINK, i.e., a link," is embedded is given to these character strings, The record of the HTML filter table which consists of the character string and classification, and URL (Uniform Resource Locator) of the link destination which is the position information on the link destination of each character string is generated for every character string of the.

[0038]In S106 which the record of the HTML filter table 122 is specified one [ at a time ] in order in the word extraction part 123, and continues in S105, It is distinguished by the word extraction part 123 whether the data in which the classification of the character string shown in the specified record is shown is either "STRING" or "LINK." And if the result of this distinction becomes in Yes, logging of the word which constitutes the character string shown in that record in S107 will be performed by the word extraction part 123. And in S108 continuing, the started word is made into a title, and the index which matched with the word of the title the title which is the page in which the word was contained, and position information is generated by the index registering part 124, and is registered into the index file 310.

[0039]On the other hand, if the result of the discrimination processing of S105 becomes in No, processing will progress to S109. In S109, it is distinguished by the word extraction part 123 whether the specification of S105 mentioned above about all the records of the HTML filter table 122 was made, and if the result of this distinction becomes in Yes, processing will progress to S110. On the other hand, if the result of this discrimination processing becomes in No, the processing which processing returned and mentioned above to S104 will be repeated.

[0040]In S111 which the record of the HTML filter table 122 is anew specified one [ at a time ] in order, and continues by the word extraction part 123 S110, It is distinguished by the word extraction part 123 whether the data in which the classification of the character string shown in the specified record is shown is "LINK." And if the result of this distinction becomes in Yes, logging of the word which constitutes the character string shown in that record in S112 will be performed by the word extraction part 123. And the data which made the group position information on the Web page of the link destination of the character string which is started word's logging origin, and its character string in S113 continuing, The word is registered by the index registering part 124 to the record in the index file 310 which is an entry, and in S114 continuing, the link flag about the data combines and the index registering part 124 registers with the record.

[0041]On the other hand, if the result of the discrimination processing of S111 becomes in No, processing will progress to S115. In S115, it is distinguished whether the specification of S110 mentioned above about all the records of the HTML filter table 122 was made, and if the result of this distinction becomes in Yes, this index production processing will be completed. On the other hand, if the result of this discrimination processing becomes in No, the processing which processing returned and mentioned above to S110 will be repeated.

[0042]Processing to the above is index production processing. Next, the processing performed by applying to S115 from S105 is further explained using the example of drawing 3. Drawing 7 shows the data structure of the index file 310 generated by the information database Management Department 300 by index production processing which was mentioned above in the case of the example of drawing 3. In drawing 7, since it will become complicated if URL is shown as position information instead, the name of the HP1-1 grade given to each Web page shown in drawing 3 is shown.

[0043]In the following explanation, the HTML filter file about HP1-1 shown in drawing 6 by processing to S104 mentioned above shall be generated. In drawing 6, first, since the classification about the character string "it is Welcome! to the Hakone hotel" of this record is "STRING" when a top record is specified by processing of S105, the discriminated result of S106 serves as Yes, and processing progresses to S107.

[0044]In S107, logging of a word is performed from a character string "it is Welcome! to the Hakone hotel." May adopt a well-known method as processing of logging of a word, for example, what is called a morphological analysis is used, The method made into the word which acquired the canonical form of the word for the part of speech and conjugated form of the word which were started using various kinds of dictionaries, and started the word of the canonical form from the character string, What is called an N gram method that starts the word of length N mechanically in order may be adopted shifting logging of a character string of one character at a time from the head of the character string.

[0045]Here, "Hakone" and a "hotel" should be started as a word from the character string "it is Welcome! to the Hakone hotel." At the title of the Web page which made the entry each of the word "Hakone" started by processing of the front step, and the "hotel" in S108 and from which the word was extracted, i.e., here, "The Hakone hotel", The index which made "H.P.1-1" the group is generated and it registers with the index file 310 at the position information on this Web page, i.e., here. By this processing of S108, each data of each data of "Hakone" of the 1st line of the index file shown in drawing 7, "H.P.1-1", and the "Hakone hotel" and the "hotel" of the 2nd line, "H.P.1-1", and the "Hakone hotel" is registered.

[0046]Next, although the result of the discrimination processing of S109 serves as No and the record of the 2nd line of a HTML filter file is specified by processing of S105, since the classification of this record is "BR", the result of the discrimination processing of S106 serves as No. Then, the result of the discrimination processing of S109 serves as No, and the record of the 3rd line of a HTML filter file is specified by processing of S106. Since the classification of the character string "accommodation plan" of this record is "LINK", the result of the discrimination processing of S106 serves as Yes, and processing progresses to S107.

[0047]In S107, logging of a character string "accommodation plan" to a character string is performed, and a word "stay" and a "plan" are started. At the title of the Web page which made the entry these word "stay" and "plans" of each and from which that word was extracted in S108, i.e., here, "The Hakone hotel", The index which made "H.P.1-1" the group is generated

and it registers with the index file 310 at the position information on this Web page, i.e., here. By this processing of S108, each data of each data of the 3rd-line "stay" of the index file shown in drawing 7, "H.P.1-1", and the "Hakone hotel" and the "plan" of the 4th line, "H.P.1-1", and the "Hakone hotel" is registered.

[0048]Processing with the same said of a character string "circumference sightseeing guidance" and a "traffic & map" is performed hereafter, If each data from the 1st row about the entry "circumference" applied to the 9th line from the 5th line of the index file shown in drawing 7, "sightseeing", "guidance", "traffic", and a "map" to the 3rd row is registered, the result of the discrimination processing of S109 will serve as No, and processing will progress to S110.

[0049]Next, although the record of the 1st line of a HTML filter file is anew specified by processing of S110, since the classification of this record is "STRING", the result of the discrimination processing of S111 serves as No, and processing progresses to S115. Since the result of the discrimination processing of S115 serves as No, processing returns to S105, and the record of the 2nd line of a HTML filter file is specified by this processing of S105 here, but since the classification of this record is "BR", the result of the discrimination processing of S106 serves as No again.

[0050]Then, the result of the discrimination processing of S115 serves as No, and the record of the 3rd line of a HTML filter file is specified by processing of S110. Since the classification of the character string "accommodation plan" of this record is "LINK", the result of the discrimination processing of S111 serves as Yes, and processing progresses to S112.

[0051]In S112, logging of a character string "accommodation plan" to a character string is performed, and a word "stay" and a "plan" are started. The data which made the group position information "H.P.1-2" on the Web page of the link destination of the character string "accommodation plan" which is word "stay" and "plan's" which were started logging origin in S113, and its character string, The word is registered by the index registering part 124 to the record in the index file 310 which is an entry, and in S114 continuing, the link flag about the data combines and the index registering part 124 registers with the record. The 3rd-line "stay" of the index file shown in drawing 7 by this processing of S113, Each data of each data of "H.P.1-2" and an "accommodation plan" and the "plan" of the 4th line, "H.P.1-2", and an "accommodation plan" is registered, and "1" is registered into the link flag of the 7th row of each of the 3rd line and the 4th line of an index file by processing of Scontinuing 114.

[0052]Processing with the same said of a character string "circumference sightseeing guidance" and a "traffic & map" is performed hereafter, If each data from the 5th row about the entry "circumference" applied to the 9th line from the 5th line of the index file shown in drawing 7, "sightseeing", "guidance", "traffic", and a "map" to the 7th row is registered, The result of the discrimination processing of S115 serves as Yes, and the index production processing about HP1-1 is completed.

[0053]The index file shown in drawing 7 is a thing when index production processing about HP2-1 is performed following the index production processing about HP1-1. When the index

record already registered when registering an index into an index file further as shown in the figure is investigated first and the same entry exists, it is made to perform additional registration to the index record about the already registered same entry, without newly generating a record.

[0054]Next, an example is shown and explained about the details of processing of the collection of a Web page performed at the information retrieval Management Department 200 which the information retrieval site 1 has, and generation of an index. Drawing 8 is a flow chart which shows the contents of processing of the retrieval processing performed by the information management retrieval part 200. First, the contents of processing of retrieval processing are explained along with the figures.

[0055]In S201, only when it is distinguished by the information retrieval section 210 whether the search formula in which the word which are the demand of information retrieval and the conditions of that search is shown has been sent from the browser 30 and this discriminated result is set to Yes, processing progresses to S202. In the information retrieval section 210, if a search formula is sent, the search formula is stored in the search formula storage 211.

[0056]In S202, the sent search formula is analyzed by the information retrieval section 210, and the word which is a search condition is started from the search formula. Search of the entry of the index file 310 which used the search word as the key is performed by the information retrieval section 210 in the turn that the search word was started S203.

[0057]In S204, as a result of search at a front step, it is distinguished by the information retrieval section 210 whether the entry which is in agreement with a search word was discovered, if the result of this distinction becomes in Yes, processing will progress to S205, and if No becomes, processing will progress to S207. In S206 which all of position information, a character string, and a link flag are acquired from the retrieval record in which the entry which is in agreement with a search word was contained by the information retrieval section 210, and continues by it in S205, The record which consists of this acquired entry, and position information, a character string and a link flag is stored in the search-results file 320 by the information retrieval section 210.

[0058]It is distinguished [ which was mentioned above about all the search words started in S207 by the processing of S202 mentioned above ] by the information retrieval section 210 whether processing of search of S203 was performed, If the result of this distinction becomes in Yes, processing will progress to S208, if the result of this distinction becomes on the other hand in No, processing will return to S203 and processing mentioned above about the search word in which processing of search is not yet performed will be performed.

[0059]The position information applicable to all the search words started by the processing of S202 mentioned above among the position information stored in the search-results file 320 in S208 here, That is, the position information included common to all the records of the search-results file 320 is extracted from the search-results file 320 by the search-results Management Department 220.

[0060]In S209. [ whether processing at a front step was able to extract position information, and ] That is, it is distinguished by the search-results Management Department 220 whether the position information included common to all the records of the search-results file 320 existed, if the result of this distinction becomes in Yes, processing will progress to S210, and if No becomes, processing will progress to S213.
[0061]In S210, the position information extracted by the processing of S208 mentioned above and the character string which is matched and is stored in the position information in the search-results file 320 are stored in the search-results list file 330 by the search-results Management Department 220.
[0062]The number of the link flag which is matched and is stored in the position information extracted by the processing of S208 mentioned above in the search-results file 320 in S211 is calculated by the search-results Management Department 220 for every position information of the, A counting result is stored in the search-results list file 330.
[0063]In S212, the search-results list file 330 is sorted by the search-results Management Department 220 so that it may become descending of the enumerated data of the link flag calculated by the front step. In S214 which the HTML file which expresses the contents of the search-results list after sorting by a Web page in S213 based on the search-results list file 330 is created by the HTML preparing part 221, and follows, The search-results Management Department 220 sends out the HTML file addressed and created to the browser 30 which is the transmitting origin of the search formula mentioned above in the WWW server part 400, and this retrieval processing ends it.
[0064]Processing to the above is retrieval processing. Next, the case where what was shown in drawing 7 as the index file 310 is stored in the information database Management Department 300 is made into an example, and this retrieval processing is explained. First, if a search formula is sent from the browser 30, the result of distinction of S201 will serve as Yes, and logging of a search word will be performed in S202 continuing. Here, each word of "Hakone", "sightseeing", and "guidance" should be started as a search word as a result of processing of S202.
[0065]If logging of a search word is completed, processing will progress to S203, first, search of a search word "Hakone" is performed about the entry of the index file 310, and the record about the entry "Hakone" in drawing 7 is discovered. Therefore, the result of the discrimination processing of S204 serves as Yes, and processing follows it to S205.
[0066]In S205, all of position information, a character string, and a link flag are acquired from the discovered record, and the record which consists of position information, a character string, and a link flag in S206 continuing is stored in the search-results file 320. Then, although discrimination processing in S207 is performed, since processing of search of S203 is not yet performed about "guidance" among the search words "sightseeing" started by the processing of S202 mentioned above, the result of the discrimination processing of S207 serves as No, and processing returns to S203.

[0067]Henceforth, the same processing as the search word "Hakone" mentioned above about a search word "sightseeing" and "guidance" is performed, The record which an entry "sightseeing" and the record about "guidance" are discovered, and consists of the position information, character string, and the link flag and the search word "Hakone" in the record is stored in the search-results file 320 from the index file 310 shown in drawing 7.
[0068]The contents of the search-results file 320 generated by processing to the above are shown in drawing 9. After the search-results file 320 shown in this drawing 9 is generated, the result of the discrimination processing of S207 serves as Yes, and processing progresses to S208.
[0069]In S208, three, HP1-1, HP2-1, and HP2-2, are extracted as position information which extraction of the position information included common to all the records of the search-results file 320 is performed, and is included common to all the records of "Hakone", "sightseeing", and "guidance" as a result. Therefore, the result of the discrimination processing of Scontinuing 209 serves as Yes, and processing progresses to S210.
[0070]The character string which is matched and is stored in the position information in S210 in position information HP1-1 extracted, HP2-1 and HP2-2, and the search-results file 320 is stored in the search-results list file 330, In S211 continuing, the number of the link flag which is matched and is stored in each of position information HP1-1 extracted, HP2-1, and HP2-2 is calculated, respectively, and the counting result is stored in the search-results list file 330.
[0071]Drawing 10 is explained here. The figure shows the contents of the search-results list file 330, and what is shown in the figure (a) is created as the search-results list file 330 by the processing to S211 mentioned above. Since the link flag about position information HP1-1 and HP2-1 is not stored in the search-results file 320 shown in drawing 9 at all, The link flag number is set to "0" about position information HP1-1 and HP2-1 in the search-results list file 330 shown in drawing 10 (a). On the other hand, since a total of three link flags are stored about position information HP2-2 in the search-results file 320 shown in drawing 9, the link flag number about position information HP2-2 in the search-results list file 330 shown in drawing 10 (a) is set to "3."
[0072]Two character strings, "circumference sightseeing guidance" and the "Hakone peripheral guide", are stored as a character string about position information HP2-2 in the search-results list file 330 shown in drawing 10 (a). Thus, when the character strings stored in the same position information in the search-results file 320 shown in drawing 9 differ, all the different character strings are stored in the search-results list file 330.
[0073]In S212 continuing, creation of the search-results list file 330 which shows drawing 10 (a) the contents by processing to S211 mentioned above will perform sorting of the search-results list file 330 so that it may become descending of the enumerated data of a link flag. The result to which sorting based on the enumerated data of a link flag was carried out to the search-results list file 330 of drawing 10 (a) is shown in drawing 10 (b).
[0074]Then, in S213, the HTML file which the HTML file which expresses the contents of the

search-results list file 330 to which sorting was performed like drawing 10 (b) by a Web page was created, and was created in S214 continuing is sent out, and this retrieval processing is completed.

[0075]The example of a screen of the Web page which shows the result of information retrieval displayed when created HTML is perused by the browser 30 is shown in drawing 11. In the screen shown in drawing 11, to the character string of "circumference sightseeing guidance" and the "Hakone peripheral guide." The link of HP2-2 by which position information is matched with those character strings in the search-results list file 330 shown in drawing 10 (b) is embedded, Similarly, the link of HP1-1 is embedded at a character string "Hakone hotel", and the link of HP2-1 is embedded at the character string "Hakone tourist agency", respectively. Thus, in creation processing of an HTML file [ in / in the HTML preparing part 221 / S213 ], The link to the Web page by which position information is matched with each character string by the search-results list file 330 creates the HTML file currently embedded at the display of those character strings.

[0076]The control program for making it carry out to the computer which has the standard composition which mentioned above the index production processing and retrieval processing which the information site 1 was performing in the embodiment of explained this invention by the above, and the same processing is created, By making the control program read into the computer, and performing it, this invention can be carried out by such computer.

[0077]It is also possible to carry out this invention by computer by making such a control program record on the recording medium which can be read by computer, making the program read from a recording medium to a computer, and performing it. The example of the possible recording medium of reading the control program made to record by computer is shown in drawing 12. The memory storage 502, such as ROM with which the computer 501 is equipped as built-in or external attachment as a recording medium, for example as shown in the figure, and a hard disk drive, or a flexible disk, Portable good signifier recording-media 503 grades, such as MO (magneto-optical disc), CD-ROM, and DVD-ROM, can be used. A recording medium may be the memory storage 506 which is connected with the computer 501 via the network 504 and with which the computer which functions as the program server 505 is provided. In this case, the transmission signal acquired by modulating a subcarrier with the data signal expressing a control program, The control program concerned can be executed now by making it transmit through the network 504 which is a transmission medium from the program server 5055, restoring to the transmission signal received by computer 501, and reproducing a control program.

---

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

---

[Brief Description of the Drawings]
[Drawing 1]The information retrieval site which carries out this invention is an entire configuration **** figure of the communication network which provides an information search service.
[Drawing 2]It is a figure showing the detailed composition of the Research and Data Processing Department, the information retrieval Management Department, and the information database Management Department.
[Drawing 3]It is a figure showing the example of the Web page collected by the information retrieval site.
[Drawing 4]It is a figure showing the HTML sauce of HP1-1 in drawing 3.
[Drawing 5]It is a flow chart which shows the contents of processing of index production processing.
[Drawing 6]It is a figure showing the HTML filter table generated from HP1-1 of drawing 3.
[Drawing 7]It is a figure explaining the data structure of the index file which is generated in the case of the example of drawing 3.
[Drawing 8]It is a flow chart which shows the contents of processing of retrieval processing.
[Drawing 9]It is a figure showing the example of a search-results file.
[Drawing 10]It is a figure showing the situation of sorting of a search-results list file.
[Drawing 11]It is a figure showing the example of a screen of the Web page which shows the result of information retrieval.
[Drawing 12]It is a figure showing the example of the possible recording medium of reading the control program made to record by computer.
[Description of Notations]
1 Information retrieval site
2a, 2b, and 2c and 2d Offer-of-information site
3a, 3b user terminal
4 Internet

20, 20a, 20b, 20c, 20d Web page
30, 30a, and 30b Browser
100 Research and Data Processing Department
110 Web page collecting part
120 Index build part
121 Web page analyzing parts
122 HTML filter table
123 Word extraction part
124 Index registering part
200 Information retrieval Management Department
210 Information retrieval section
211 Search formula storage
220 Search-results Management Department
221 HTML preparing part
300 Information database Management Department
310 Index file
320 Search-results file
330 Search-results list file
400 WWW server part
501 Computer
502 and 506 Memory storage
503 Portable good signifier recording media
504 Network
505 Program server

---

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

---

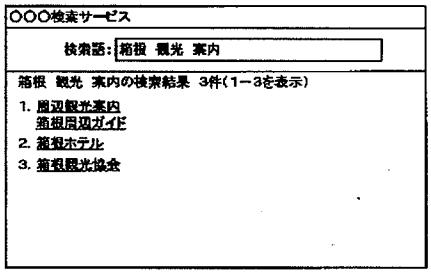
[Drawing 6]

図3のHP1－1から生成されるHTMLフィルタテーブルを示す図

| 種別 | 文字列 | 位置情報(リンク先URL) |
|---|---|---|
| STRING | 箱根ホテルへようこそ | |
| BR | | |
| LINK | 宿泊プラン | http://www.aaaa_hotel.co.jp/shikuhaku.html |
| BR | | |
| LINK | 周辺観光案内 | http://www.xxxx.or.jp/hakone.html |
| BR | | |
| LINK | 交通＆マップ | http://www.aaaa_hotel.co.jp/access.html |
| BR | | |
| EOF | | |

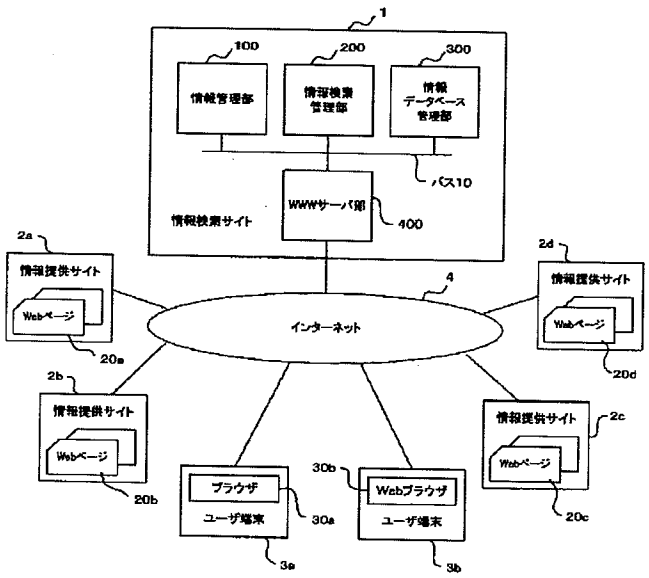
[Drawing 11]

情報検索の結果を示すWebページの画面例を示す図



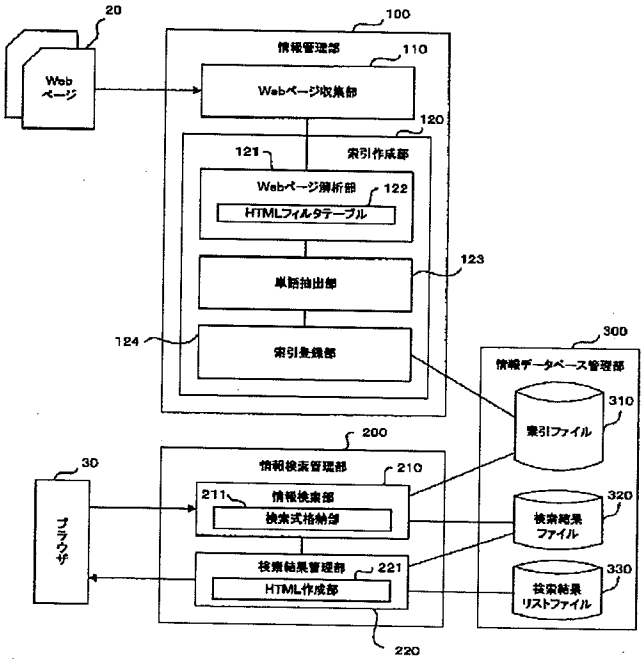

[Drawing 1]

本発明を実施する情報検索サイトが情報検索サービスを提供する通信ネットワークの全体構成を示す図




[Drawing 2]

情報管理部、情報検索管理部、及び情報データベース管理部の詳細構成を示す図




[Drawing 4]

図3におけるHP1－1のHTMLソースを示す図.

(a)ブラウザ表示画面

箱根ホテル

箱根ホテルへようこそ

宿泊プラン

周辺観光案内

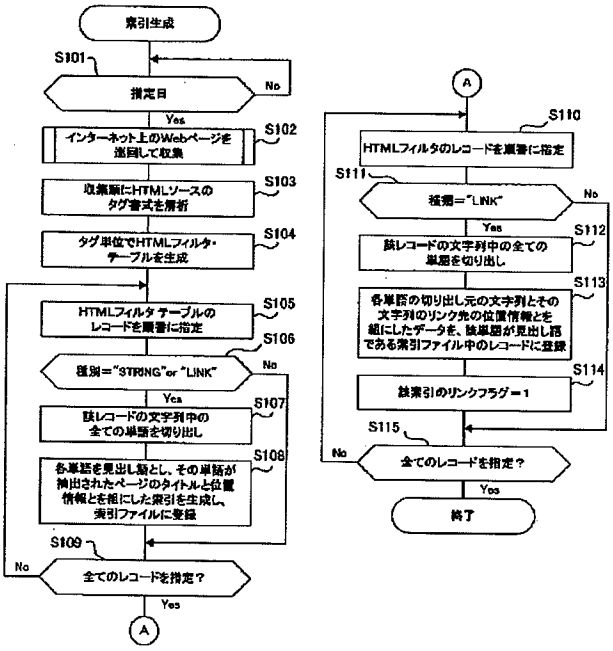
交通&マップ

(b)HTMLソース

```
<HTML>
<HEAD>
 <TITLE>箱根ホテル</TITLE>
</HEAD>
<BODY>
<FONT SIZE="+2">箱根ホテルへようこそ</FONT><BR>
<P> </P>
<P> </P>
<A HREF="http://www.aaaa_hotel.co.jp/shikuhaku.html">宿泊プラン</A><BR>
<P> </P>
<A HREF="http://www.xxxx.or.jp/hakone.html>周辺観光案内</A><BR>
<P> </P>
<A HREF="http://www.aaaa_hotel.co.jp/access.html">交通&マップ</A><BR>
</BODY>
</HTML>
```

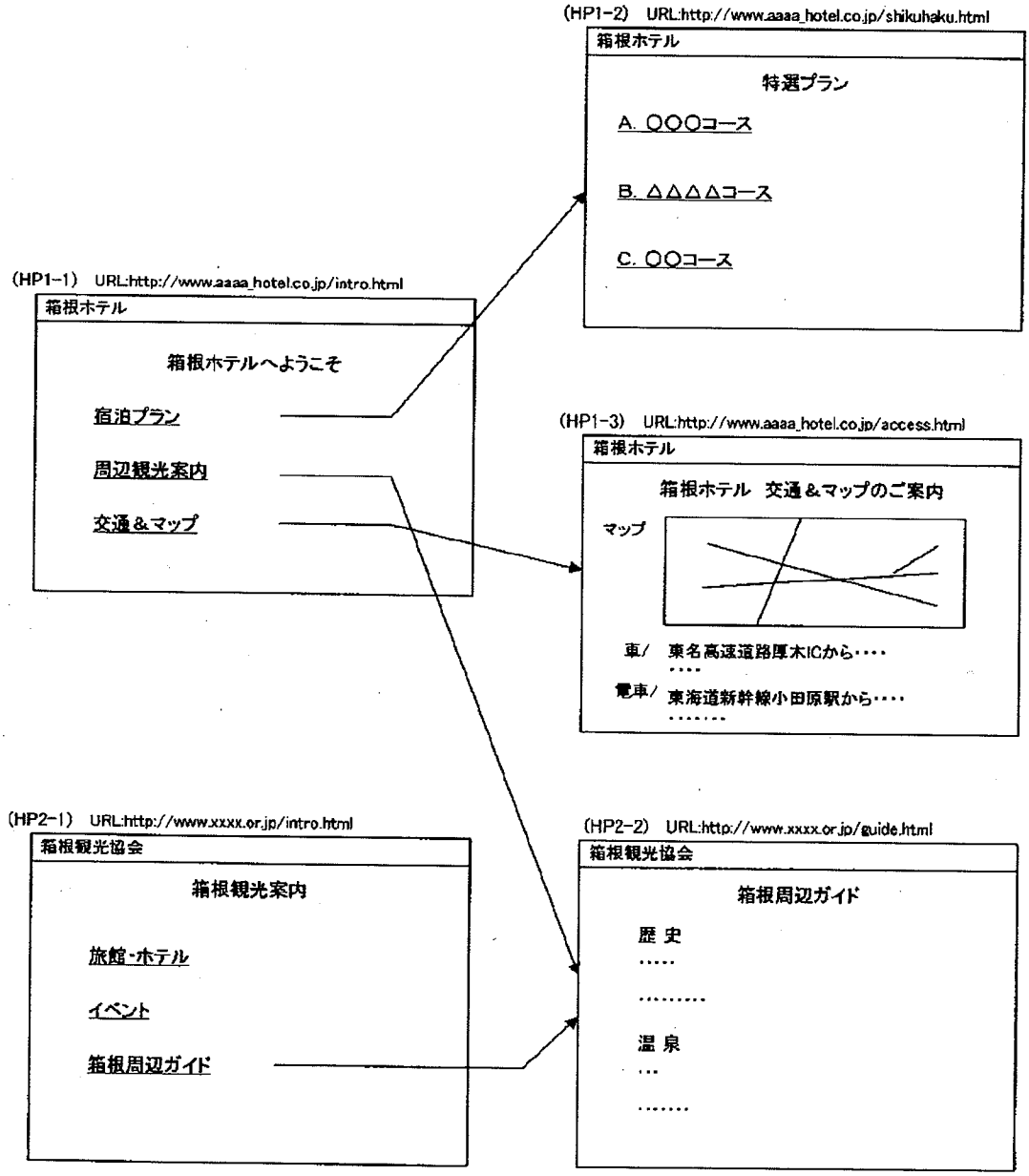
## [Drawing 5]

索引生成処理の処理内容を示すフローチャート



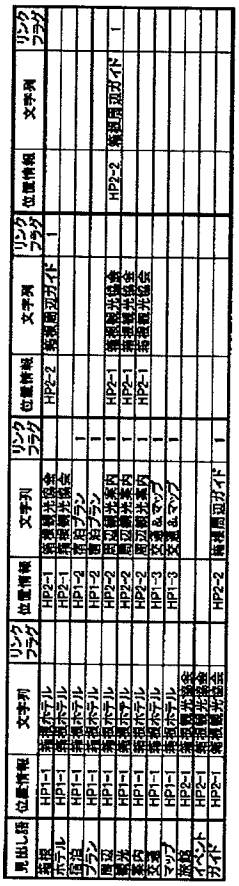

## [Drawing 3]

情報検索サイトによって収集されるWebページの例を示す図




[Drawing 7]

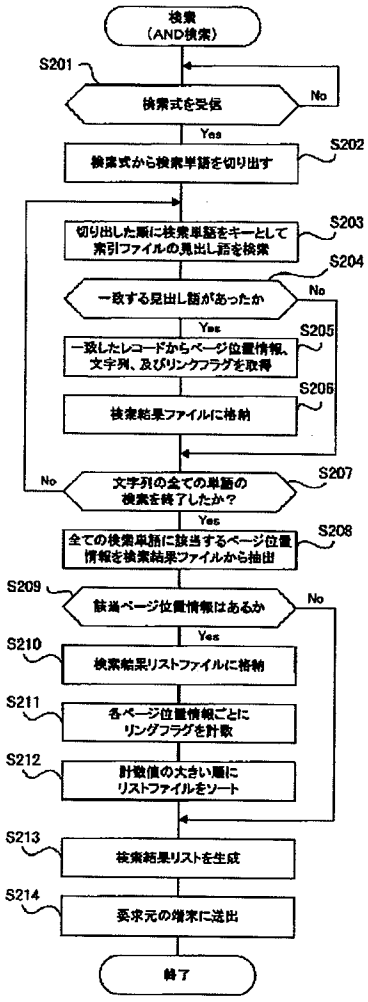
図3の例の場合に生成される索引ファイルのデータ構造を説明する図



| 見出し語 | 位置情報 | 文字列 | リンクフラグ | 位置情報 | 文字列 | リンクフラグ | 位置情報 | 文字列 | リンクフラグ |
|---|---|---|---|---|---|---|---|---|---|
| 浅草 | HP1-1 | 浅草ホテル | | HP2-1 | 浅草観光協会 | | HP2-2 | 浅草周辺ガイド | 1 |
| ホテル | HP1-1 | 浅草ホテル | | HP2-1 | 浅草観光協会 | | | | |
| 宿泊 | HP1-1 | 浅草ホテル | | HP1-4 | 宿泊プラン | 1 | | | |
| プラン | HP1-1 | 浅草ホテル | | HP1-2 | 宿泊プラン | 1 | | | |
| 周辺 | HP1-1 | 浅草ホテル | | HP2-2 | 周辺観光案内 | 1 | HP2-1 | 浅草観光協会 | |
| 観光 | HP1-1 | 浅草ホテル | | HP2-2 | 周辺観光案内 | 1 | HP2-1 | 浅草観光協会 | |
| 案内 | HP1-1 | 浅草ホテル | | HP2-2 | 周辺観光案内 | 1 | HP2-1 | 浅草観光協会 | |
| 交通 | HP1-1 | 浅草ホテル | | HP1-3 | 交通&マップ | 1 | | | |
| マップ | HP1-1 | 浅草ホテル | | HP1-3 | 交通&マップ | 1 | | | |
| 周辺 | HP2-1 | 浅草観光協会 | | | | | | | |
| イベント | HP2-1 | 浅草観光協会 | | | | | | | |
| ガイド | HP2-1 | 浅草観光協会 | | HP2-2 | 浅草周辺ガイド | 1 | | | |

[Drawing 8]

検索処理の処理内容を示すフローチャート




[Drawing 9]

**検索結果ファイルの例を示す図**

| 見出し語 | 位置情報 | 文字列 | リンクフラグ | 位置情報 | 文字列 | リンクフラグ | 位置情報 | 文字列 | リンクフラグ |
|---|---|---|---|---|---|---|---|---|---|
| 箱根 | HP1-1 | 箱根ホテル | | HP2-1 | 箱根観光協会 | | HP2-2 | 箱根周辺ガイド | 1 |
| 観光 | HP1-1 | 箱根ホテル | | HP2-2 | 周辺観光案内 | 1 | HP2-1 | 箱根観光協会 | |
| 案内 | HP1-1 | 箱根ホテル | | HP2-2 | 周辺観光案内 | 1 | HP2-1 | 箱根観光協会 | |

[Drawing 10]

**検索結果リストファイルのソートの様子を示す図**

(a)ソート前

| 位置情報 | 文字列 | リンクフラグ数 | 位置情報 | 文字列 | リンクフラグ数 | 位置情報 | 文字列 | リンクフラグ数 |
|---|---|---|---|---|---|---|---|---|
| HP1-1 | 箱根ホテル | 0 | HP2-1 | 箱根観光協会 | 0 | HP2-2 | 周辺観光案内<br>箱根周辺ガイド | 3 |

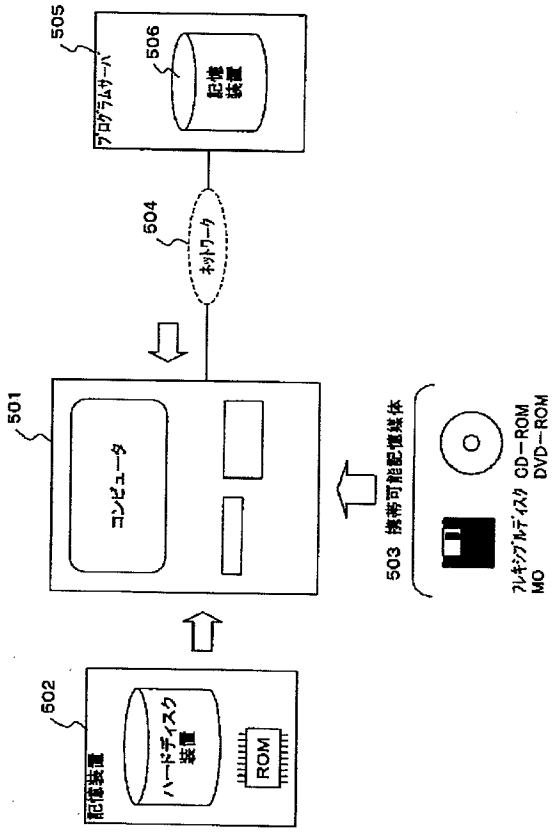
ソート(降順:リンクフラグ数)

(b)ソート後

| 位置情報 | 文字列 | リンクフラグ数 | 位置情報 | 文字列 | リンクフラグ数 | 位置情報 | 文字列 | リンクフラグ数 |
|---|---|---|---|---|---|---|---|---|
| HP2-2 | 周辺観光案内<br>箱根周辺ガイド | 3 | HP1-1 | 箱根ホテル | 0 | HP2-1 | 箱根観光協会 | 0 |

[Drawing 12]

記憶させた制御プログラムをコンピュータで
読み取ることの可能な記録媒体の例を示す図




[Translation done.]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2003－30235
（P2003－30235A）
(43)公開日　平成15年１月31日(2003.1.31)

| (51)Int.Cl.[7] | 識別記号 | ＦＩ | ﾃｰﾏｺｰﾄﾞ(参考) |
|---|---|---|---|
| Ｇ０６Ｆ　17/30 | ３４０ | Ｇ０６Ｆ　17/30 | ３４０Ｚ　５Ｂ０７５ |
| | ４１９ | | ４１９Ｂ |

審査請求　未請求　請求項の数６　ＯＬ　（全 14 頁）

(21)出願番号　特願2001－212184(P2001－212184)

(22)出願日　平成13年７月12日(2001.7.12)

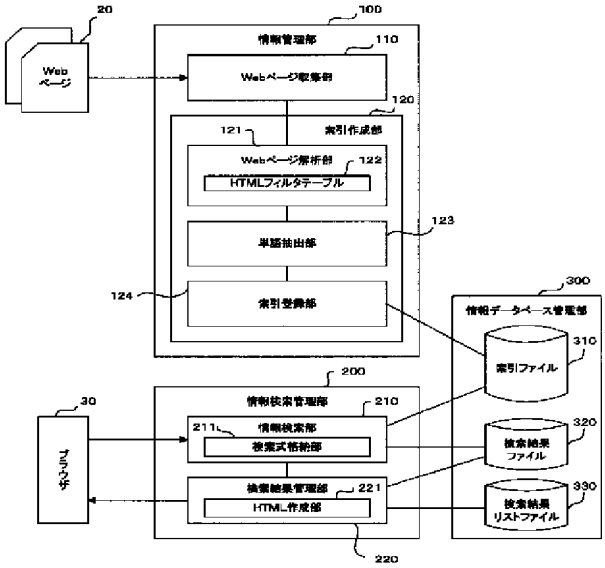
(71)出願人　000001443
カシオ計算機株式会社
東京都渋谷区本町１丁目６番２号
(72)発明者　寺田　俊仁
東京都羽村市栄町３丁目２番１号　カシオ計算機株式会社羽村技術センター内
(74)代理人　100093632
弁理士　阪本　紀康

Ｆターム(参考)　5B075 ND36 PQ75 PR10

(54)【発明の名称】　情報検索システム、情報検索方法、及びプログラム

(57)【要約】
【課題】　検索エンジンが提供する情報検索の結果を情報検索者の検索目的に対してより適切なものにする。
【解決手段】　単語抽出部１２３はＷｅｂページ２０に含まれている文字列から単語を抽出する。索引登録部１２４は抽出された文字列のうちハイパーリンクが埋め込まれているものについて該リンク先の位置情報を取得し、該単語に、Ｗｅｂページ２０の位置情報と該リンク先の位置情報とを対応付けて索引ファイル３１０に登録する。情報検索部２１０は索引ファイル３１０を検索し、検索対象を表す単語に対応付けられている位置情報を取得して検索結果ファイル３２０に格納する。検索結果管理部２２０は検索結果ファイル３２０の情報をソートして多くのハイパーリンクが張られているリンク先の位置情報に高い優先順位を与えたものを表現するＨＴＭＬファイルを作成してブラウザ３０に提供する。

情報管理部、情報検索管理部、及び情報データベース管理部の詳細構成を示す図



【特許請求の範囲】
【請求項１】　通信ネットワーク上で公開されている文書情報に含まれている単語と、該通信ネットワーク上の論理的な位置を示す情報であって該単語に関連する情報が存在する文書情報位置を示す単語関連位置情報とを対応付けてなる索引ファイルを用意し、検索の対象を表す単語に基づいて該索引ファイルを検索して該検索対象を表す単語に対応している単語関連位置情報を提示するシステムであって、
前記文書情報に含まれている文字列から単語を抽出する抽出手段と、
前記文字列のうち、該文字列に関連する情報が提供されている参照先についての前記位置を示す情報である参照先位置情報を有していることを示す属性が付されているものについて、該参照先位置情報を取得する参照先位置取得手段と、
前記抽出手段によって抽出された単語を、該単語の抽出元の文字列が含まれていた文書情報についての前記位置を示す文書位置情報と前記属性が付されている該文字列についての前記参照先位置情報とからなる単語関連位置情報に対応付けて前記索引ファイルに登録する登録手段と、
前記検索対象を表す単語に基づいて前記索引ファイルの検索を行なって該単語に対応付けられている単語関連位置情報を該索引ファイルから取得する検索手段と、
前記検索手段によって取得された単語関連位置情報のうち、前記参照先位置情報を優先して提示する提示手段と、
を有することを特徴とする情報検索システム。
【請求項２】　前記提示手段は、前記検索手段によって取得された単語関連位置情報に同一の前記位置を示す前記単語関連位置情報と参照先位置情報とが含まれているときには、該単語関連位置情報のうち参照先位置情報として取得された数の多いものが優先されるようにして該単語関連位置情報を提示することを特徴とする請求項１に記載の情報検索システム。
【請求項３】　前記文書情報は、Ｗｅｂページを表現するページ記述言語によって記述されており、
前記参照先位置取得手段は、前記文字列に埋め込まれていたハイパーリンクにおけるリンク先の前記位置を示す情報を前記参照先位置情報として取得する、
ことを特徴とする請求項１に記載の情報検索システム。
【請求項４】　前記登録手段は、前記索引ファイルに、前記文書位置情報によって前記位置が示されている文書情報に付されているタイトルである文字列に該文書位置情報を対応付ける登録、及び、前記参照先位置情報によって前記位置が示されている前記リンク先へのハイパーリンクが埋め込まれている文字列に該参照先情報を対応付ける登録を行ない、
前記提示手段は、前記索引ファイルに登録されている対応付けに基づいたハイパーリンクの埋め込まれた文字列であって前記単語関連位置情報で示される前記位置へのリンクを示す該文字列を用いて該単語関連位置情報を提示する、
ことを特徴とする請求項３に記載の情報検索システム。
【請求項５】　通信ネットワーク上で公開されている文書情報に含まれている単語と、該通信ネットワーク上の論理的な位置を示す情報であって該単語に関連する情報が存在する文書情報位置を示す単語関連位置情報とを対応付けてなる索引ファイルを用意し、検索の対象を表す単語に基づいて該索引ファイルを検索して該検索対象を表す単語に対応している単語関連位置情報を提示する方法であって、
前記文書情報に含まれている文字列から単語の抽出を行ない、
前記文字列のうち、該文字列に関連する情報が提供されている参照先についての前記位置を示す情報である参照先位置情報を有していることを示す属性が付されているものについて、該参照先位置情報を取得し、
前記抽出によって抽出された単語を、該単語の抽出元の文字列が含まれていた文書情報についての前記位置を示す文書位置情報と前記属性が付されている該文字列についての前記参照先位置情報とからなる単語関連位置情報に対応付けて前記索引ファイルに登録し、
前記検索対象を表す単語に基づいて前記索引ファイルの検索を行なって該単語に対応付けられている単語関連位置情報を該索引ファイルから取得し、
前記検索によって取得された単語関連位置情報のうち、前記参照先位置情報を優先して提示する、
ことを特徴とする情報検索方法。
【請求項６】　コンピュータに実行させることにより、通信ネットワーク上で公開されている文書情報に含まれている単語と、該通信ネットワーク上の論理的な位置を示す情報であって該単語に関連する情報が存在する文書情報位置を示す単語関連位置情報とを対応付けてなる索引ファイルを用意し、検索の対象を表す単語に基づいて該索引ファイルを検索して該条件を表す単語に対応している単語関連位置情報を提示する処理を該コンピュータに行なわせるためのプログラムであって、
前記文書情報に含まれている文字列から単語の抽出を行なう処理と、
前記文字列のうち、該文字列に関連する情報が提供されている参照先についての前記位置を示す情報である参照先位置情報を有していることを示す属性が付されているものについて、該参照先位置情報を取得する処理と、
前記抽出によって抽出された単語を、該単語の抽出元の文字列が含まれていた文書情報についての前記位置を示す文書位置情報と前記属性が付されている該文字列についての前記参照先位置情報とからなる単語関連位置情報に対応付けて前記索引ファイルに登録する処理と、

前記検索対象を表す単語に基づいて前記索引ファイルの検索を行なって該単語に対応付けられている単語関連位置情報を該索引ファイルから取得する処理と、
前記検索によって取得された単語関連位置情報のうち、前記参照先位置情報を優先して提示する処理と、
をコンピュータに行なわせるためのプログラム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、情報を検索する技術に関し、特に、検索の要求に対し、その要求により適切に合致した情報を提供できるようにする技術に関する。
【０００２】
【従来の技術】近年、インターネットの普及により、インターネット上のWWW（World WideWeb）システムで提供されているＷｅｂページの数は爆発的に増え続けている。また、インターネット上では、この膨大な情報の中から目的とする情報を検索するサービスを提供する検索エンジンが多数開設されている。
【０００３】検索エンジンがネット上の情報を収集する方式のひとつとして、ロボット型と称されているものがある。ロボット型の検索エンジンでは、スパイダあるいはクローラなどと呼ばれるロボットプログラムが定期的に起動されて、インターネット上で公開されているＷｅｂページを表現しているＨＴＭＬ（HyperText MarkupLanguage ）ファイルの自動収集が行なわれる。情報検索が行なわれるときには、検索エンジンを利用する情報検索者が目的とする情報に関係の深いキーワードを検索サイトに与えることにより、収集されたファイルからそのキーワードが含まれたものを抽出する処理が行なわれ、そのキーワードの含まれているＷｅｂページのリストが、そのＷｅｂページについてのインターネット上における論理的な位置を示す情報と共に、検索結果として情報検索者に提供される。
【０００４】
【発明が解決しようとする課題】一般に、ロボット型の検索エンジンは、情報の収集から検索結果の提供に至るまでの全ての処理をコンピュータで自動的に行なっており、人間の判断による情報の操作は介在しないので、収集された情報の属するジャンルやその情報の質についての整理がなされていない。そのため、情報の検索の際に、単なるキーワードの一致による検索を行なっていたのでは、重要な情報を含むＷｅｂページが検索結果に埋もれてしまったり、あるいは、いわゆる検索ノイズ、すなわち有用性の低い情報しか含まれていないＷｅｂページばかり検索結果に多く含まれてしまったりする場合が少なくなかった。
【０００５】以上の問題を鑑み、検索エンジンが提供する情報検索の結果を情報検索者の検索目的に対してより適切なものにすることが本発明が解決しようとする課題である。
【０００６】
【課題を解決するための手段】本発明は、通信ネットワーク上で公開されている文書情報に含まれている単語と、該通信ネットワーク上の論理的な位置を示す情報であって該単語に関連する情報が存在する文書情報位置を示す単語関連位置情報とを対応付けてなる索引ファイルを用意し、検索の対象を表す単語に基づいて該索引ファイルを検索して該検索対象を表す単語に対応している単語関連位置情報を提示するシステムまたは方法を前提とする。
【０００７】そして、本発明の態様のひとつである情報検索システムでは、前記文書情報に含まれている文字列から単語を抽出する抽出手段と、前記文字列のうち、該文字列に関連する情報が提供されている参照先についての前記位置を示す情報である参照先位置情報を有していることを示す属性が付されているものについて、該参照先位置情報を取得する参照先位置取得手段と、前記抽出手段によって抽出された単語を、該単語の抽出元の文字列が含まれていた文書情報についての前記位置を示す文書位置情報と前記属性が付されている該文字列についての前記参照先位置情報とからなる単語関連位置情報に対応付けて前記索引ファイルに登録する登録手段と、前記検索対象を表す単語に基づいて前記索引ファイルの検索を行なって該単語に対応付けられている単語関連位置情報を該索引ファイルから取得する検索手段と、前記検索手段によって取得された単語関連位置情報のうち、前記参照先位置情報を優先して提示する提示手段とを有するように構成することによって前述した課題を解決する。
【０００８】例えば技術論文において、他の多くの論文によって参照されている論文は特に重要度が高いものと考えることができる。本発明はこの考え方に基づくものであり、他の文書情報から参照されている情報を示す位置情報はそうでないものと比べてより適切なものであるとみなす。すなわち、単語関連位置情報において参照先位置情報は単なる文書位置情報よりも重要であるとみなし、単語関連位置情報提示手段に参照先位置情報を優先的に提示させるようにするのである。こうすることによって、通信ネットワーク上で公開されている文書情報における文字列であって検索条件に合致する単語が含まれている該文字列が参照している参照先を示す参照先位置情報が単なる文書位置情報よりも優先的に提示されるので、情報検索の結果が情報検索者の検索目的に対してより適切なものとなる。
【０００９】なお、上述した本発明に係る情報検索システムにおいて、前記提示手段は、前記検索手段によって取得された単語関連位置情報に同一の前記位置を示す前記単語関連位置情報と参照先位置情報とが含まれているときには、該単語関連位置情報のうち参照先位置情報として取得された数の多いものが優先されるようにして該

単語関連位置情報を提示するようにしてもよい。
【００１０】この構成によれば、他の文字列から参照される数のより多い情報の位置が示されている単語関連位置情報ほど優先されて提示されるようになる。また、前述した本発明に係る情報検索システムにおいて、前記文書情報は、Ｗｅｂページを表現するページ記述言語によって記述されており、前記参照先位置取得手段は、前記文字列に埋め込まれていたハイパーリンクにおけるリンク先の前記位置を示す情報を前記参照先位置情報として取得するようにしてもよく、この構成によっても本発明に係る情報検索システムと同様の作用・効果を奏する。
【００１１】なお、このとき、前記登録手段は、前記索引ファイルに、前記文書位置情報によって前記位置が示されている文書情報に付されているタイトルである文字列に該文書位置情報を対応付ける登録、及び、前記参照先位置情報によって前記位置が示されている前記リンク先へのハイパーリンクが埋め込まれている文字列に該参照先情報を対応付ける登録を行ない、前記提示手段は、前記索引ファイルに登録されている対応付けに基づいたハイパーリンクの埋め込まれた文字列であって前記単語関連位置情報で示される前記位置へのリンクを示す該文字列を用いて該単語関連位置情報を提示するようにしてもよい。
【００１２】この構成によれば、検索結果として提示される単語関連位置情報が文字列に埋め込まれたハイパーリンクとして提示されるので、その検索結果を受け取った情報検索者はリンク先へ簡単に辿り着くことが可能となり、検索条件の単語に関連する情報を容易に取得できるようになる。
【００１３】また、本発明の態様のひとつである情報検索方法は、前記文書情報に含まれている文字列から単語の抽出を行ない、前記文字列のうち、該文字列に関連する情報が提供されている参照先についての前記位置を示す情報である参照先位置情報を有していることを示す属性が付されているものについて、該参照先位置情報を取得し、前記抽出によって抽出された単語を、該単語の抽出元の文字列が含まれていた文書情報についての前記位置を示す文書位置情報と前記属性が付されている該文字列についての前記参照先位置情報とからなる単語関連位置情報に対応付けて前記索引ファイルに登録し、前記検索対象を表す単語に基づいて前記索引ファイルの検索を行なって該単語に対応付けられている単語関連位置情報を該索引ファイルから取得し、前記検索によって取得された単語関連位置情報うち、前記参照先位置情報を優先して提示することにより、前述した本発明に係る情報検索システムと同様の作用・効果が得られる。
【００１４】なお、上述した本発明に係る情報検索方法と同様の手順からなる処理をコンピュータに行なわせるためのプログラムでも、そのプログラムをコンピュータに実行させることによって前述した課題を解決することができる。
【００１５】
【発明の実施の形態】以下、本発明の実施の形態を図面に基づいて説明する。図１は本発明を実施する情報検索サイトが情報検索サービスを提供する通信ネットワークの全体構成を示す図である。
【００１６】図１において、情報検索サイト１、情報提供サイト２ａ、２ｂ、２ｃ、２ｄ、及びユーザ端末３ａ、３ｂはいずれも通信ネットワークであるインターネット４に接続されており、相互にデータの授受を行なうことができる。情報検索サイト１は、ユーザ端末３ａ及びａｂにロボット検索型の情報検索サービスを提供するＷＷＷサーバシステムであり、情報管理部１００、情報検索管理部２００、情報データベース管理部３００、及びＷＷＷサーバ管理部４００を備えて構成されている。
【００１７】情報管理部１００は、インターネット４上に公開されている情報の自動収集を行ない、収集された情報を情報データベース管理部３００に蓄積する。情報検索管理部２００は、インターネット４を介して送られてくる情報検索の要求に応じて情報データベース管理部３００に蓄積されている情報の検索を行ない、その検索の結果を要求元に返送する。
【００１８】情報データベース管理部３００では、情報管理部１００によって収集された情報の蓄積、及び情報検索管理部２００による情報の検索が行なわれる。ＷＷＷサーバ部４００は、インターネット４を介して送られてくる収集された情報を情報管理部１００に転送する処理、インターネット４を介して送られてくる情報検索の要求を情報検索管理部２００に転送する処理、及び情報検索管理部２００から送られてくる情報検索の結果を示す情報が表されているＷｅｂページの送出の処理が行なわれる。
【００１９】情報提供サイト２ａ、２ｂ、２ｃ、及び２ｄは、それぞれＷｅｂページ２０ａ、２０ｂ、２０ｃ、及び２０ｄをインターネット４上で公開するＷＷＷサーバシステムである。なお、図１においては４つの情報提供サイトを示しているが、インターネット４に接続される情報提供サイトの数は任意でよい。
【００２０】ユーザ端末３ａ及び３ｂは、それぞれ情報提供サイト２ａ、２ｂ、２ｃ、及び２ｄや情報検索サイト１から提供されるＷｅｂページを閲覧するソフトウェアであるブラウザ３０ａ及び３０ｂを実行可能なコンピュータであり、インターネット４上で公開されている情報の検索を情報検索サイト１へ依頼する情報検索者によって操作される。なお、図１においては２つのユーザを示しているが、インターネット４に接続されるユーザ端末の数も任意でよい。
【００２１】なお、これらの情報検索サイト１、情報提供サイト２ａ、２ｂ、２ｃ、及び２ｄ、ユーザ端末３ａ及び３ｂは、いずれも標準的な構成を有するコンピュー

タ、すなわち、制御プログラムを実行することで各構成要素を制御するＣＰＵと、ＲＯＭやＲＡＭ及び磁気記憶装置などからなり、ＣＰＵに各構成要素を制御させる制御プログラムの記憶やＣＰＵが制御プログラムを実行する際のワークエリアあるいは各種データの記憶領域として使用される記憶部と、ユーザによる操作に対応する各種のデータが取得される入力部と、ディスプレイなどに各種のデータを提示してユーザに通知する出力部と、ネットワークに接続するためのインタフェース機能を提供するＩ／Ｆ部とを備えるコンピュータを用いて構成することもできる。

【００２２】次に図２について説明する。同図は、図１における情報検索サイト１に備えられている情報管理部１００、情報検索管理部２００、及び情報データベース管理部３００の構成を更に詳細に示したものである。図２に示すように、情報管理部１００はＷｅｂページ収集部１１０及び索引生成部１２０を備えて構成されており、情報管理検索部２００は情報検索部２１０及び検索結果管理部２２０を備えて構成されており、情報データベース管理部３００は索引ファイル３１０、検索結果ファイル３２０、及び検索結果リストファイル３３０を備えて構成されている。

【００２３】Ｗｅｂページ収集部１１０は、インターネット４上で公開されているＷｅｂページ２０の収集を行なう。索引作成部１２０は、Ｗｅｂページ収集部１１０によって収集されたＷｅｂページ２０の位置情報、すなわちＷｅｂページ２０が存在するインターネット４上の論理的な位置を示す位置情報を引くことのできる索引を索引ファイル３１０に登録する。索引作成部１２０はＷｅｂページ解析部１２１、単語抽出部１２３、索引登録部１２４を備えて構成されている。

【００２４】Ｗｅｂページ解析部１２１はＷｅｂページ２０の解析を行なってＷｅｂページ２０を表現しているＨＴＭＬファイルの本文に記述されている各ＨＴＭＬタグをレコードの単位とするＨＴＭＬフィルタテーブル１２２を作成する。単語抽出部１２３は、ＨＴＭＬフィルタテーブル１２２に示されている文字列から単語の抽出を行なう。

【００２５】索引登録部１２４は、単語抽出部１２３によって抽出された単語とＷｅｂページ２０についての位置情報との関係、及び、Ｗｅｂページ２０でその単語にハイパーリンク（以下、単に「リンク」と略す）が埋め込まれているときにはリンクの存在及びその単語とそのリンク先であるＷｅｂページの位置情報との関係を示す索引データを索引ファイル３１０に登録する。

【００２６】情報検索部２１０は、インターネット４に接続されているいずれかのユーザ端末で実行されているブラウザ３０の制御によってそのユーザ端末から送られてくる情報検索の要求をＷＷＷサーバ部４００から取得して、その情報検索の条件を示す検索式を取り出して検索式格納部２１１に格納する。そして、索引ファイル３００を検索してその検索式に示されている単語（キーワード）が見出しとなっている索引データを取得して検索結果ファイル３２０に格納する。

【００２７】検索結果管理部２２０は、情報検索部２１０による検索が完了すると、検索結果ファイル３２０に格納されている索引データに示されている位置情報と、その位置情報へ張られているリンクの数の合計とを検索結果リストファイル３３０に格納する。そして、検索結果リストファイル３３０に格納された位置情報をそのリンク数に従ってソートし、ソートされた位置情報ならなる検索結果リストが表示されるＷｅｂページを表現するＨＴＭＬファイルをＨＴＭＬ作成部２２１で作成する。作成されたＨＴＭＬファイルはブラウザ３０が実行されているユーザ端末へ宛ててＷＷＷサーバ部４００によりインターネット４に送出される。

【００２８】次に、情報検索サイト１の有する情報管理部１００において行なわれる、Ｗｅｂページの収集及び索引の生成の処理の詳細について、具体例を提示して説明する。図３は、インターネット４に公開されていて情報検索サイト１によって収集されるＷｅｂページ２０の例を示している。

【００２９】図３にはＨＰ１－１、ＨＰ１－２、ＨＰ１－３、ＨＰ２－１、ＨＰ２－２の計５枚のＷｅｂページが例示されている。なお、同図に示されている矢印はリンクの関係を示している。すなわち、例えばＨＰ１－１の「宿泊プラン」なる文字列にはＨＰ１－２へのリンクが埋め込まれていることを示している。

【００３０】また、図４にはＨＰ１－１のＨＴＭＬソースが示されている。同図（ｂ）に示すＨＴＭＬがブラウザ３０によって閲覧されると同図（ａ）に示す画面が表示される。ここで図５について説明する。同図は情報管理部１００で実行される索引生成処理の処理内容を示すフローチャートである。この処理が実行されることによって、Ｗｅｂページの収集及び索引の生成が情報管理部１００で行なわれる。

【００３１】まず、Ｓ１０１において、現在の日付が、予め指定されているＷｅｂページ２０の収集指定日であるか否かが判別され、この判定結果がＹｅｓ、すなわち現在がその指定日になったときにのみ、処理がＳ１０２に進む。この日付の指定の仕方は任意であるが、例えば毎月の月末最終日などといった指定が行なわれる。

【００３２】Ｓ１０２ではＷｅｂページ収集部１１０によるインターネット４上で公開されているＷｅｂページ２０の巡回・収集の処理が行なわれる。この巡回・収集の手法は周知のロボット型検索エンジンで従来から行なわれているものをそのまま利用すればよい。

【００３３】Ｓ１０３では、収集されたＷｅｂページ２０のＨＴＭＬソースのタグ書式がＷｅｂページ解析部１２１によって解析され、続くＳ１０４においてＨＴＭＬ

フィルタテーブルがＷｅｂページ解析部１２１によって生成される。図３に示したＨＰ１－１から生成されるＨＴＭＬフィルタテーブルを図６に示す。Ｗｅｂページ解析部１２１では、図４（ｂ）に示したＨＰ１－１についてのＨＴＭＬソースが解析されて図６に示すＨＴＭＬフィルタテーブルが生成される。

【００３４】Ｓ１０３の処理内容を図４（ｂ）を参照しながら更に説明すると、Ｗｅｂページ解析部１２１では、解析対象のＨＴＭＬソースの本文、すなわち<;BODY>;の開始タグと終了タグとの間に挟まれている全ての文字列が<;BR>;タグ（改行タグ）を文字列の区切りとみなされて抽出される。

【００３５】続くＳ１０４の処理では、選択された文字列、その文字列が選択された理由を示す表示、及び他のＷｅｂページへのリンクが埋め込まれているものについてはそのリンク先の位置情報が１つのレコードとして纏められてＨＴＭＬフィルタテーブル１２２が生成される。

【００３６】図４（ｂ）に示すＨＴＭＬソースから図６に示すＨＴＭＬフィルタテーブルが作成される様子について説明すると、まず、ＨＴＭＬソースにおける本文の記述部分である<;BODY>;タグの開始タグと終了タグとの間に挟まれている部分、すなわち<;BODY>;タグと<;/BODY>; タグとに挟まれている部分に含まれている文字列は、<;BR>;タグによって「箱根ホテルへようこそ」、「宿泊プラン」、「周辺観光案内」、「交通＆マップ」なる４つの文字列に区切られている。

【００３７】そして、これらの文字列のうち、「箱根ホテルにようこそ」には、リンクが埋め込まれていない文字列であることを示す「ＳＴＲＩＮＧ」という種別が与えられてＨＴＭＬフィルタテーブルのレコードがひとつ生成される。また、「宿泊プラン」、「周辺観光案内」、「交通＆マップ」の各文字列はいずれも他のＷｅｂページへのリンクが埋め込まれているので、これらの文字列には「ＬＩＮＫ」、すなわちリンクが埋め込まれている文字列であるという種別が与えられ、その文字列と種別と各文字列のリンク先の位置情報であるそのリンク先のＵＲＬ（Uniform Resource Locator）とからなるＨＴＭＬフィルタテーブルのレコードがその文字列毎に生成される。

【００３８】Ｓ１０５では、単語抽出部１２３においてＨＴＭＬフィルタテーブル１２２のレコードが順番にひとつずつ指定され、続くＳ１０６において、その指定されたレコードに示されている文字列の種別を示すデータが「ＳＴＲＩＮＧ」若しくは「ＬＩＮＫ」のいずれかであるか否かが単語抽出部１２３によって判別される。そして、この判別の結果がＹｅｓならば、Ｓ１０７においてそのレコードに示されている文字列を構成している単語の切り出しが単語抽出部１２３で行なわれる。そして続くＳ１０８において、切り出された単語を見出しとし、その単語が含まれていたページのタイトルと位置情報とをその見出しの単語に対応付けた索引が索引登録部１２４で生成されて索引ファイル３１０に登録される。

【００３９】一方、Ｓ１０５の判別処理の結果がＮｏならばＳ１０９に処理が進む。Ｓ１０９では、ＨＴＭＬフィルタテーブル１２２の全てのレコードについて前述したＳ１０５の指定がなされたか否かが単語抽出部１２３で判別され、この判別の結果がＹｅｓならばＳ１１０に処理が進む。一方、この判別処理の結果がＮｏならばＳ１０４へ処理が戻って上述した処理が繰り返される。

【００４０】Ｓ１１０では、単語抽出部１２３でＨＴＭＬフィルタテーブル１２２のレコードが改めて順番にひとつずつ指定され、続くＳ１１１において、その指定されたレコードに示されている文字列の種別を示すデータが「ＬＩＮＫ」であるか否かが単語抽出部１２３によって判別される。そして、この判別の結果がＹｅｓならば、Ｓ１１２においてそのレコードに示されている文字列を構成している単語の切り出しが単語抽出部１２３で行なわれる。そして続くＳ１１３において、切り出された単語の切り出し元である文字列とその文字列のリンク先のＷｅｂページの位置情報とを組にしたデータが、その単語が見出し語である索引ファイル３１０中のレコードへ索引登録部１２４によって登録されると共に、続くＳ１１４において索引登録部１２４によってそのデータについてのリンクフラグが併せてそのレコードに登録される。

【００４１】一方、Ｓ１１１の判別処理の結果がＮｏならばＳ１１５に処理が進む。Ｓ１１５では、ＨＴＭＬフィルタテーブル１２２の全てのレコードについて前述したＳ１１０の指定がなされたか否かが判別され、この判別の結果がＹｅｓならばこの索引生成処理が終了する。一方、この判別処理の結果がＮｏならばＳ１１０へ処理が戻って上述した処理が繰り返される。

【００４２】以上までの処理が索引生成処理である。次に、Ｓ１０５からＳ１１５にかけて行なわれる処理を図３の例を用いて更に説明する。図７は、図３の例の場合に上述した索引生成処理によって情報データベース管理部３００に生成される索引ファイル３１０のデータ構造を示している。なお、図７においては、位置情報としてＵＲＬを示すと煩雑になるので、その代わりに図３に示した各Ｗｅｂページに付したＨＰ１－１等の名称を示している。

【００４３】なお、以下の説明においては、前述したＳ１０４までの処理によって図６に示したＨＰ１－１についてのＨＴＭＬフィルタファイルが生成されているものとする。図６において、まず、先頭のレコードがＳ１０５の処理によって指定されると、このレコードの文字列「箱根ホテルへようこそ」についての種別は「ＳＴＲＩＮＧ」であるので、Ｓ１０６の判別結果はＹｅｓとなり、処理はＳ１０７に進む。

【００４４】Ｓ１０７では、文字列「箱根ホテルへようこそ」から単語の切り出しが行なわれる。なお、単語の切り出しの処理には周知の方式を採用してよく、例えばいわゆる形態素解析を利用し、切り出した単語の品詞と活用形を各種の辞書を用いてその単語の標準形を取得してその標準形の単語を文字列から切り出した単語とする方式や、文字列の切り出しをその文字列の先頭から１文字ずつずらしながら順に長さＮの語を機械的に切り出すいわゆるＮグラム方式を採用してもよい。

【００４５】ここでは、文字列「箱根ホテルへようこそ」から「箱根」及び「ホテル」が単語として切り出されたものとする。Ｓ１０８では、前ステップの処理によって切り出された単語「箱根」及び「ホテル」の各々を見出し語とし、その単語が抽出されたＷｅｂページのタイトル、すなわちここでは「箱根ホテル」と、このＷｅｂページの位置情報、すなわちここでは「ＨＰ１－１」とを組にした索引が生成され、索引ファイル３１０に登録される。このＳ１０８の処理によって、図７に示す索引ファイルの第１行目の「箱根」、「ＨＰ１－１」、「箱根ホテル」の各データ、及び第２行目の「ホテル」、「ＨＰ１－１」、「箱根ホテル」の各データが登録される。

【００４６】次にＳ１０９の判別処理の結果がＮｏとなり、Ｓ１０５の処理によってＨＴＭＬフィルタファイルの第２行目のレコードが指定されるが、このレコードの種別は「ＢＲ」なのでＳ１０６の判別処理の結果はＮｏとなる。続いてＳ１０９の判別処理の結果がＮｏとなり、Ｓ１０６の処理によってＨＴＭＬフィルタファイルの第３行目のレコードが指定される。このレコードの文字列「宿泊プラン」の種別は「ＬＩＮＫ」なのでＳ１０６の判別処理の結果はＹｅｓとなり、処理はＳ１０７に進む。

【００４７】Ｓ１０７では、文字列「宿泊プラン」から文字列の切り出しが行なわれ、単語「宿泊」及び「プラン」が切り出される。Ｓ１０８では、この単語「宿泊」及び「プラン」各々を見出し語とし、その単語が抽出されたＷｅｂページのタイトル、すなわちここでは「箱根ホテル」と、このＷｅｂページの位置情報、すなわちここでは「ＨＰ１－１」とを組にした索引が生成され、索引ファイル３１０に登録される。このＳ１０８の処理によって、図７に示す索引ファイルの第３行目の「宿泊」、「ＨＰ１－１」、「箱根ホテル」の各データ、及び第４行目の「プラン」、「ＨＰ１－１」、「箱根ホテル」の各データが登録される。

【００４８】以下、文字列「周辺観光案内」及び「交通＆マップ」についても同様の処理が施され、図７に示す索引ファイルの第５行目から第９行目にかけての見出し語「周辺」、「観光」、「案内」、「交通」、「マップ」についての第１列目から第３列目までの各データが登録されると、Ｓ１０９の判別処理の結果がＮｏとなり、処理はＳ１１０に進む。

【００４９】次に、Ｓ１１０の処理によってＨＴＭＬフィルタファイルの第１行目のレコードが改めて指定されるが、このレコードの種別は「ＳＴＲＩＮＧ」なのでＳ１１１の判別処理の結果はＮｏとなり、処理はＳ１１５に進む。ここではＳ１１５の判別処理の結果はＮｏとなるので、処理はＳ１０５へと戻り、このＳ１０５の処理によってＨＴＭＬフィルタファイルの第２行目のレコードが指定されるが、このレコードの種別は「ＢＲ」なのでＳ１０６の判別処理の結果は再びＮｏとなる。

【００５０】続いてＳ１１５の判別処理の結果がＮｏとなり、Ｓ１１０の処理によってＨＴＭＬフィルタファイルの第３行目のレコードが指定される。このレコードの文字列「宿泊プラン」の種別は「ＬＩＮＫ」なのでＳ１１１の判別処理の結果はＹｅｓとなり、処理はＳ１１２に進む。

【００５１】Ｓ１１２では、文字列「宿泊プラン」から文字列の切り出しが行なわれ、単語「宿泊」及び「プラン」が切り出される。Ｓ１１３では、切り出された単語「宿泊」及び「プラン」の切り出し元である文字列「宿泊プラン」とその文字列のリンク先のＷｅｂページの位置情報「ＨＰ１－２」とを組にしたデータが、その単語が見出し語である索引ファイル３１０中のレコードへ索引登録部１２４によって登録されると共に、続くＳ１１４において索引登録部１２４によってそのデータについてのリンクフラグが併せてそのレコードに登録される。このＳ１１３の処理によって、図７に示す索引ファイルの第３行目の「宿泊」、「ＨＰ１－２」、「宿泊プラン」の各データ、及び第４行目の「プラン」、「ＨＰ１－２」、「宿泊プラン」の各データが登録され、続くＳ１１４の処理によって索引ファイルの第３行目及び第４行目のそれぞれ第７列目のリンクフラグに「１」が登録される。

【００５２】以下、文字列「周辺観光案内」及び「交通＆マップ」についても同様の処理が施され、図７に示す索引ファイルの第５行目から第９行目にかけての見出し語「周辺」、「観光」、「案内」、「交通」、「マップ」についての第５列目から第７列目までの各データが登録されると、Ｓ１１５の判別処理の結果がＹｅｓとなり、ＨＰ１－１についての索引生成処理が終了する。

【００５３】なお、図７に示した索引ファイルは、ＨＰ１－１についての索引生成処理に続き、ＨＰ２－１についての索引生成処理が行なわれたときのものである。同図のように、索引ファイルに更に索引を登録するときには、既に登録されている索引レコードをまず調べ、同一の見出し語が存在するときには、新たにレコードを生成せずに、既に登録されている同一の見出し語についての索引レコードに追加登録を行なうようにする。

【００５４】次に、情報検索サイト１の有する情報検索管理部２００において行なわれる、Ｗｅｂページの収集

及び索引の生成の処理の詳細について、具体例を提示して説明する。図８は情報管理検索部２００で実行される検索処理の処理内容を示すフローチャートである。まず、同図に沿って検索処理の処理内容を説明する。

【００５５】Ｓ２０１では、情報検索の要求及びその検索の条件である単語が示されている検索式がブラウザ３０から送られてきたか否かが情報検索部２１０で判別され、この判別結果がＹｅｓとなったときにのみ、処理がＳ２０２に進む。なお、情報検索部２１０では、検索式が送られてくるとその検索式を検索式格納部２１１に格納する。

【００５６】Ｓ２０２では送られてきた検索式が情報検索部２１０で解析され、その検索式から検索条件である単語が切り出される。Ｓ２０３では、検索単語が切り出された順番で、その検索単語をキーとした索引ファイル３１０の見出し語の検索が情報検索部２１０によって行なわれる。

【００５７】Ｓ２０４では、前ステップでの検索の結果、検索単語に一致する見出し語が発見されたか否かが情報検索部２１０によって判別され、この判別の結果がＹｅｓならばＳ２０５に処理が進み、ＮｏならばＳ２０７に処理が進む。Ｓ２０５では、情報検索部２１０によって、検索単語に一致する見出し語の含まれていた検索レコードから位置情報、文字列、及びリンクフラグが全て取得され、続くＳ２０６において、この取得された見出し語と、位置情報、文字列、及びリンクフラグとからなるレコードが情報検索部２１０によって検索結果ファイル３２０に格納される。

【００５８】Ｓ２０７では、前述したＳ２０２の処理によって切り出された全ての検索単語について前述したＳ２０３の検索の処理が行なわれたか否かが情報検索部２１０によって判別され、この判別の結果がＹｅｓならばＳ２０８に処理が進み、一方この判別の結果がＮｏならばＳ２０３へと処理が戻って未だ検索の処理の行なわれていない検索単語について上述した処理が行なわれる。

【００５９】ここで、Ｓ２０８において、検索結果ファイル３２０に格納されている位置情報のうち前述したＳ２０２の処理によって切り出された全ての検索単語に該当する位置情報、すなわち検索結果ファイル３２０の全てのレコードに共通に含まれている位置情報が検索結果管理部２２０によって検索結果ファイル３２０から抽出される。

【００６０】Ｓ２０９では、前ステップでの処理によって位置情報の抽出が行なえたか否か、すなわち検索結果ファイル３２０の全てのレコードに共通に含まれている位置情報が存在したか否かが検索結果管理部２２０によって判別され、この判別の結果がＹｅｓならばＳ２１０に処理が進み、ＮｏならばＳ２１３に処理が進む。

【００６１】Ｓ２１０では、前述したＳ２０８の処理によって抽出された位置情報と、検索結果ファイル３２０においてその位置情報に対応付けられて格納されている文字列とが検索結果管理部２２０によって検索結果リストファイル３３０に格納される。

【００６２】Ｓ２１１では、検索結果ファイル３２０において、前述したＳ２０８の処理によって抽出された位置情報に対応付けられて格納されているリンクフラグの個数がその位置情報毎に検索結果管理部２２０によって計数され、計数結果が検索結果リストファイル３３０に格納される。

【００６３】Ｓ２１２では、前ステップによって計数されたリンクフラグの計数値の大きい順となるように検索結果リストファイル３３０が検索結果管理部２２０によってソートされる。Ｓ２１３では、検索結果リストファイル３３０に基づき、ソートされた後の検索結果リストの内容をＷｅｂページで表現するＨＴＭＬファイルがＨＴＭＬ作成部２２１によって作成され、続くＳ２１４において、検索結果管理部２２０は、前述した検索式の送信元であるブラウザ３０へ宛てて作成されたＨＴＭＬファイルをＷＷＷサーバ部４００に送出させ、この検索処理が終了する。

【００６４】以上までの処理が検索処理である。次に、この検索処理について、索引ファイル３１０として図７に示したものが情報データベース管理部３００に格納されている場合を例にして説明する。まず、ブラウザ３０から検索式が送られてくると、Ｓ２０１の判別の結果がＹｅｓとなり、続くＳ２０２において検索単語の切り出しが行なわれる。ここでは、このＳ２０２の処理の結果、検索単語として「箱根」、「観光」、「案内」の各語が切り出されたものとする。

【００６５】検索単語の切り出しが完了すると処理はＳ２０３に進み、まず、索引ファイル３１０の見出し語について検索単語「箱根」の検索が行なわれ、図７における見出し語「箱根」についてのレコードが発見される。従ってＳ２０４の判別処理の結果はＹｅｓとなり、Ｓ２０５に処理が進む。

【００６６】Ｓ２０５では発見されたレコードから位置情報、文字列、及びリンクフラグが全て取得され、続くＳ２０６において位置情報、文字列、及びリンクフラグからなるレコードが検索結果ファイル３２０に格納される。その後、Ｓ２０７における判別処理が行なわれるが、前述したＳ２０２の処理によって切り出された検索単語のうち「観光」及び「案内」についてはＳ２０３の検索の処理が未だ行なわれていないので、Ｓ２０７の判別処理の結果はＮｏとなり、処理はＳ２０３へと戻る。

【００６７】以降、検索単語「観光」及び「案内」について上述した検索単語「箱根」と同様の処理が行なわれ、図７に示す索引ファイル３１０から見出し語「観光」及び「案内」についてのレコードが発見されてそのレコードにおける位置情報、文字列、及びリンクフラグと検索単語「箱根」とからなるレコードが検索結果ファ

イル３２０に格納される。
【００６８】以上までの処理によって生成される検索結果ファイル３２０の内容を図９に示す。この図９に示す検索結果ファイル３２０が生成された後にはＳ２０７の判別処理の結果がＹｅｓとなり、処理はＳ２０８に進む。
【００６９】Ｓ２０８では、検索結果ファイル３２０の全てのレコードに共通に含まれている位置情報の抽出が行なわれ、その結果、「箱根」、「観光」、「案内」の全てのレコードに共通に含まれている位置情報としてＨＰ１－１、ＨＰ２－１、及びＨＰ２－２の３つが抽出される。従って、続くＳ２０９の判別処理の結果はＹｅｓとなり、処理はＳ２１０に進む。
【００７０】Ｓ２１０では、抽出された位置情報ＨＰ１－１、ＨＰ２－１、及びＨＰ２－２と検索結果ファイル３２０においてその位置情報に対応付けられて格納されている文字列とが検索結果リストファイル３３０に格納され、続くＳ２１１において、抽出された位置情報ＨＰ１－１、ＨＰ２－１、及びＨＰ２－２の各々に対応付けられて格納されているリンクフラグの個数がそれぞれ計数され、その計数結果が検索結果リストファイル３３０に格納される。
【００７１】ここで図１０について説明する。同図は、検索結果リストファイル３３０の内容を示しており、上述したＳ２１１までの処理によって、同図（ａ）に示すものが検索結果リストファイル３３０として作成される。図９に示す検索結果ファイル３２０には位置情報ＨＰ１－１及びＨＰ２－１についてのリンクフラグが全く格納されていないので、図１０（ａ）に示す検索結果リストファイル３３０における位置情報ＨＰ１－１及びＨＰ２－１については、リンクフラグ数は「０」とされている。一方、図９に示す検索結果ファイル３２０における位置情報ＨＰ２－２についてはリンクフラグが合計３つ格納されているので、図１０（ａ）に示す検索結果リストファイル３３０における位置情報ＨＰ２－２についてのリンクフラグ数は「３」とされている。
【００７２】なお、図１０（ａ）に示す検索結果リストファイル３３０における位置情報ＨＰ２－２についての文字列として、「周辺観光案内」と「箱根周辺ガイド」の２つの文字列が格納されている。このように、図９に示す検索結果ファイル３２０において同一の位置情報に格納されている文字列が異なるときには、その異なる文字列の全てを検索結果リストファイル３３０に格納するようにする。
【００７３】前述したＳ２１１までの処理によって図１０（ａ）にその内容を示す検索結果リストファイル３３０が作成されると、続くＳ２１２において、リンクフラグの計数値の大きい順となるように検索結果リストファイル３３０のソートが行なわれる。図１０（ａ）の検索結果リストファイル３３０に対してリンクフラグの計数値に基づくソートの行なわれた結果が図１０（ｂ）に示されているものである。
【００７４】その後、Ｓ２１３において、図１０（ｂ）のようにソートが行なわれた検索結果リストファイル３３０の内容をＷｅｂページで表現するＨＴＭＬファイルが作成され、続くＳ２１４において作成されたＨＴＭＬファイルが送出されて、この検索処理が終了する。
【００７５】作成されたＨＴＭＬがブラウザ３０によって閲覧されることによって表示される、情報検索の結果を示すＷｅｂページの画面例を図１１に示す。図１１に示す画面において、「周辺観光案内」及び「箱根周辺ガイド」の文字列には、図１０（ｂ）に示す検索結果リストファイル３３０においてそれらの文字列に位置情報が対応付けられているＨＰ２－２へのリンクが埋め込まれており、同様に、文字列「箱根ホテル」にはＨＰ１－１へのリンクが、また、文字列「箱根観光協会」にはＨＰ２－１へのリンクがそれぞれ埋め込まれている。このように、ＨＴＭＬ作成部２２１は、Ｓ２１３におけるＨＴＭＬファイルの作成処理においては、検索結果リストファイル３３０で位置情報が各文字列に対応付けられているＷｅｂページへのリンクが、それらの文字列の表示に埋め込まれているＨＴＭＬファイルを作成する。
【００７６】なお、以上までに説明した本発明の実施形態において情報サイト１が行なっていた索引生成処理及び検索処理と同様の処理を前述したような標準的な構成を有するコンピュータに行なわせるための制御プログラムを作成し、その制御プログラムをそのコンピュータに読み込ませて実行させることにより、このようなコンピュータで本発明を実施することができる。
【００７７】また、このような制御プログラムをコンピュータで読み取り可能な記録媒体に記録させ、そのプログラムを記録媒体からコンピュータに読み出させて実行させることによって本発明をコンピュータで実施することも可能である。記録させた制御プログラムをコンピュータで読み取ることの可能な記録媒体の例を図１２に示す。同図に示すように、記録媒体としては、例えば、コンピュータ５０１に内蔵若しくは外付けの付属装置として備えられるＲＯＭやハードディスク装置などの記憶装置５０２、あるいはフレキシブルディスク、ＭＯ（光磁気ディスク）、ＣＤ－ＲＯＭ、ＤＶＤ－ＲＯＭなどといった携帯可能記録媒体５０３等が利用できる。また、記録媒体はネットワーク５０４を介してコンピュータ５０１と接続される、プログラムサーバ５０５として機能するコンピュータが備えている記憶装置５０６であってもよい。この場合には、制御プログラムを表現するデータ信号で搬送波を変調して得られる伝送信号を、プログラムサーバ５０５５から伝送媒体であるネットワーク５０４を通じて伝送するようにし、コンピュータ５０１では受信した伝送信号を復調して制御プログラムを再生することで当該制御プログラムを実行できるようになる。

【００７８】
【発明の効果】本発明によれば、索引ファイルに、通信ネットワーク上で公開されている文書情報に含まれている文字列を構成する単語を、その文字列が含まれていた文書情報の位置を示す文書位置情報、及びその文字列に関連する情報が提供されている参照先についての位置を示す参照先位置情報とからなる単語関連位置情報に対応付けて登録することで、検索対象を表す単語に基づいてその索引ファイルの検索を行った場合に、その検索によって取得された単語関連位置情報のうち、参照先位置情報を優先して提示することができるので、情報検索者の検索目的に対してより適切な情報検索の結果を提供できる。
【図面の簡単な説明】
【図１】本発明を実施する情報検索サイトが情報検索サービスを提供する通信ネットワークの全体構成示す図である。
【図２】情報管理部、情報検索管理部、及び情報データベース管理部の詳細構成を示す図である。
【図３】情報検索サイトによって収集されるＷｅｂページの例を示す図である。
【図４】図３におけるＨＰ１－１のＨＴＭＬソースを示す図である。
【図５】索引生成処理の処理内容を示すフローチャートである。
【図６】図３のＨＰ１－１から生成されるＨＴＭＬフィルタテーブルを示す図である。
【図７】図３の例の場合に生成される索引ファイルのデータ構造を説明する図である。
【図８】検索処理の処理内容を示すフローチャートである。
【図９】検索結果ファイルの例を示す図である。
【図１０】検索結果リストファイルのソートの様子を示す図である。
【図１１】情報検索の結果を示すＷｅｂページの画面例を示す図である。
【図１２】記録させた制御プログラムをコンピュータで読み取ることの可能な記録媒体の例を示す図である。
【符号の説明】
１　情報検索サイト
２ａ、２ｂ、２ｃ、２ｄ　情報提供サイト
３ａ、３ｂ　ユーザ端末
４　インターネット
２０、２０ａ、２０ｂ、２０ｃ、２０ｄ　Ｗｅｂページ
３０、３０ａ、３０ｂ　ブラウザ
１００　情報管理部
１１０　Ｗｅｂページ収集部
１２０　索引作成部
１２１　Ｗｅｂページ解析部
１２２　ＨＴＭＬフィルタテーブル
１２３　単語抽出部
１２４　索引登録部
２００　情報検索管理部
２１０　情報検索部
２１１　検索式格納部
２２０　検索結果管理部
２２１　ＨＴＭＬ作成部
３００　情報データベース管理部
３１０　索引ファイル
３２０　検索結果ファイル
３３０　検索結果リストファイル
４００　ＷＷＷサーバ部
５０１　コンピュータ
５０２、５０６　記憶装置
５０３　携帯可能記録媒体
５０４　ネットワーク
５０５　プログラムサーバ

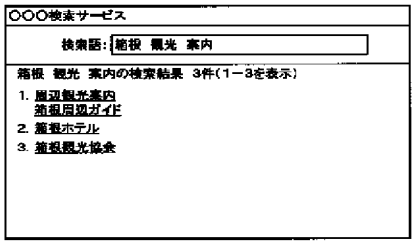
【図６】

**図3のHP1－1から生成されるHTMLフィルタテーブルを示す図**

| 種別 | 文字列 | 位置情報(リンク先URL) |
|---|---|---|
| STRING | 箱根ホテルへようこそ | |
| BR | | |
| LINK | 宿泊プラン | http://www.aaaa_hotel.co.jp/shikuhaku.html |
| BR | | |
| LINK | 周辺観光案内 | http://www.xxxx.or.jp/hakone.html |
| BR | | |
| LINK | 交通＆マップ | http://www.aaaa_hotel.co.jp/access.html |
| BR | | |
| EOF | | |

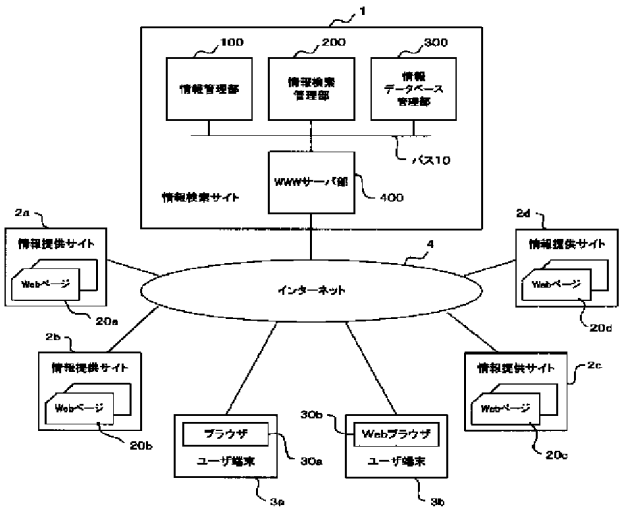
【図１１】

**情報検索の結果を示すWebページの画面例を示す図**



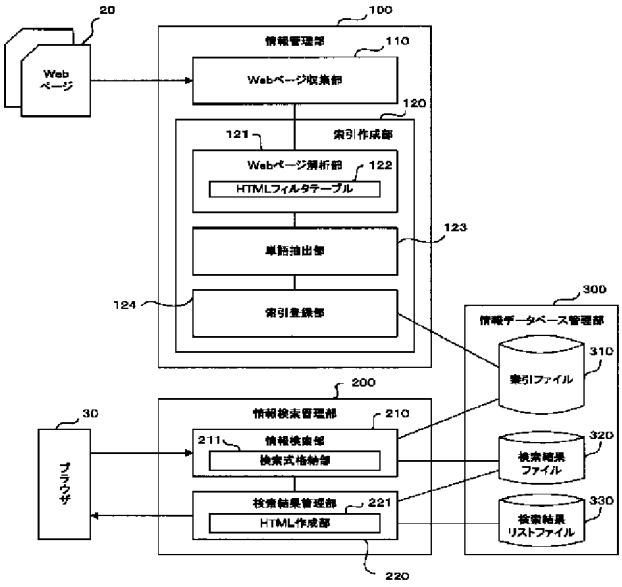
【図１】

本発明を実施する情報検索サイトが情報検索サービスを提供する通信ネットワークの全体構成を示す図




【図２】

情報管理部、情報検索管理部、及び情報データベース管理部の詳細構成を示す図




【図４】

図3におけるHP1－1のHTMLソースを示す図

（a）ブラウザ表示画面

箱根ホテル

箱根ホテルへようこそ

宿泊プラン

周辺観光案内

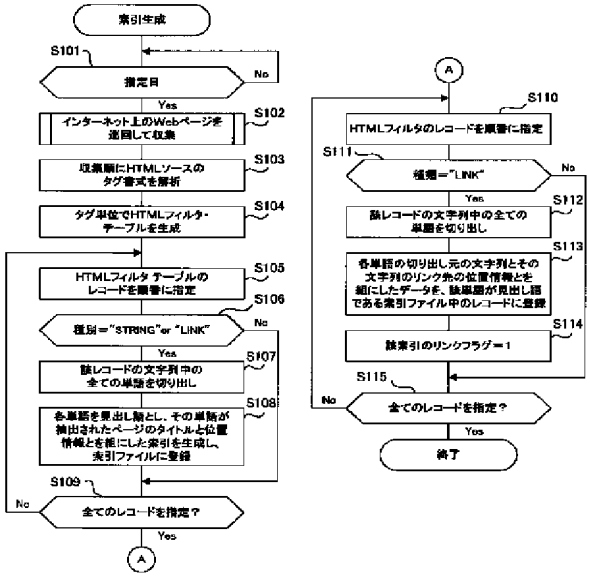
交通＆マップ

（b）HTMLソース

```
<HTML>
<HEAD>
  <TITLE>箱根ホテル</TITLE>
</HEAD>
<BODY>
<FONT SIZE="+2">箱根ホテルへようこそ</FONT><BR>
<P> </P>
<P> </P>
<A HREF="http://www.aaaa_hotel.co.jp/shikuhaku.html">宿泊プラン</A><BR>
<P> </P>
<A HREF="http://www.xxxx.or.jp/hakone.html>周辺観光案内</A><BR>
<P> </P>
<A HREF="http://www.aaaa_hotel.co.jp/access.html">交通＆マップ</A><BR>
</BODY>
</HTML>
```

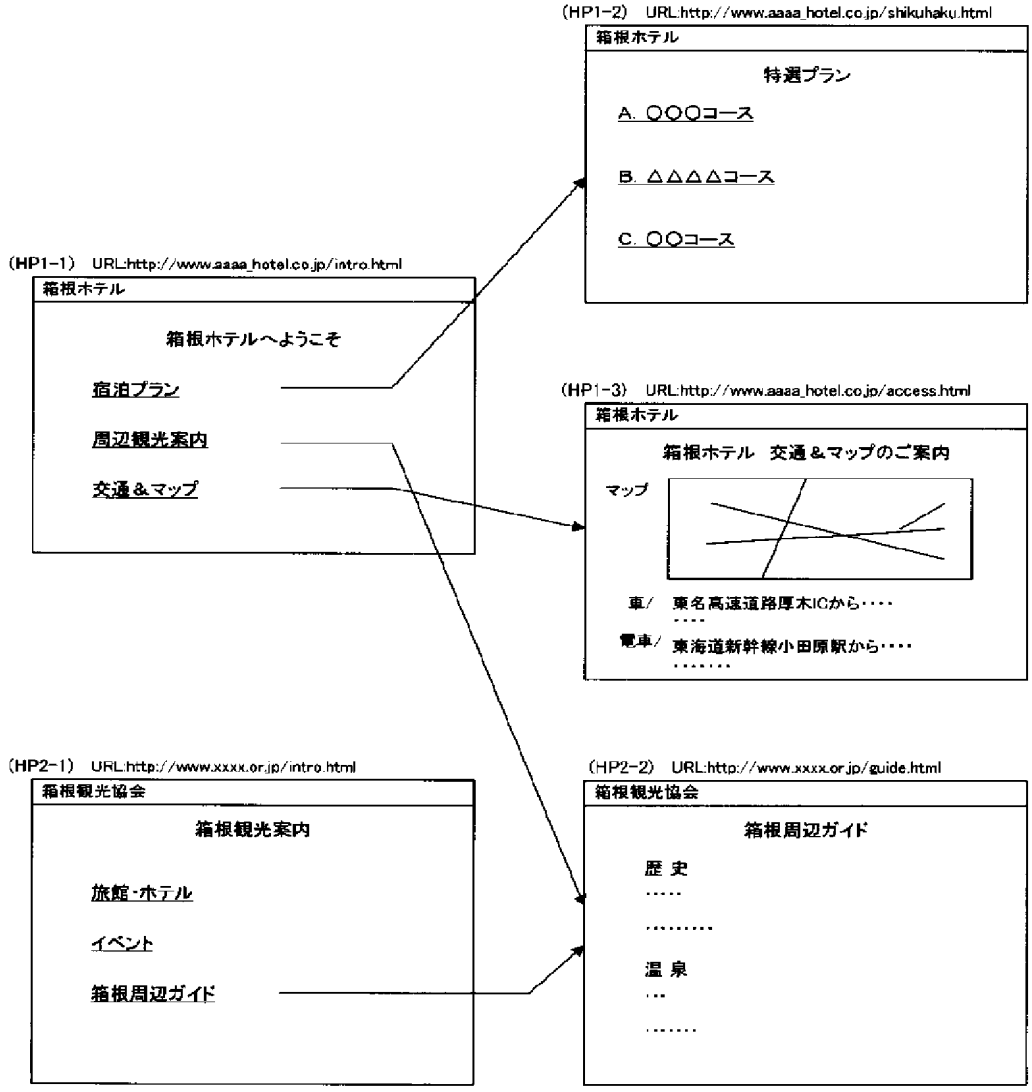
【図５】

索引生成処理の処理内容を示すフローチャート



【図３】

情報検索サイトによって収集されるWebページの例を示す図



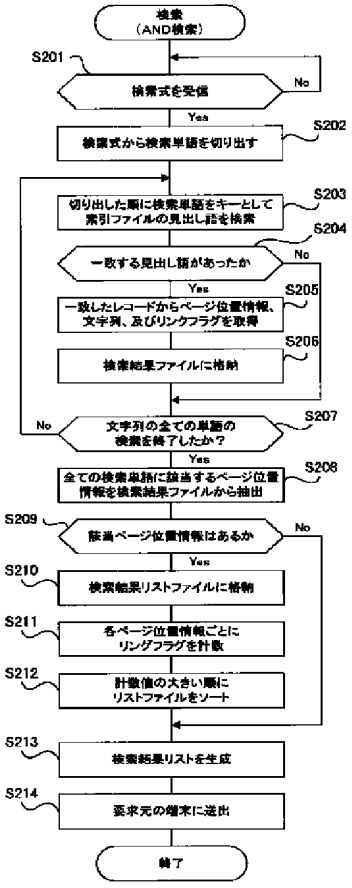
【図７】

図3の例の場合に生成される索引ファイルのデータ構造を説明する図

| 見出し語 | 位置情報 | 文字列 | リンクフラグ | 位置情報 | 文字列 | リンクフラグ | 位置情報 | 文字列 | リンクフラグ |
|---|---|---|---|---|---|---|---|---|---|
| 梅屋 | HP1-1 | 梅屋ホテル | | HP1-1 | 梅屋ホテル | | HP2-1 | 梅屋観光協会 | |
| ホテル | HP1-1 | 梅屋ホテル | | HP1-1 | 梅屋ホテル | | HP2-1 | 梅屋観光協会 | |
| 住所 | HP1-1 | 梅屋ホテル | | HP1-2 | 住所プラン | 1 | HP2-2 | 梅屋周辺ガイド | 1 |
| プラン | HP1-1 | 梅屋ホテル | | HP1-2 | 住所プラン | 1 | | | |
| 周辺 | HP1-1 | 梅屋ホテル | | HP2-2 | 周辺観光案内 | 1 | HP2-1 | 梅屋観光協会 | |
| 観光 | HP1-1 | 梅屋ホテル | | HP2-2 | 周辺観光案内 | 1 | HP2-1 | 梅屋観光協会 | |
| 案内 | HP1-1 | 梅屋ホテル | | HP2-2 | 周辺観光案内 | 1 | HP2-1 | 梅屋観光協会 | |
| 交通 | HP1-1 | 梅屋ホテル | | HP1-3 | 交通＆マップ | 1 | | | |
| マップ | HP1-1 | 梅屋ホテル | | HP1-3 | 交通＆マップ | 1 | | | |
| 近隣 | HP2-1 | 梅屋観光協会 | | | | | | | |
| イベント | HP2-1 | 梅屋観光協会 | | | | | | | |
| ガイド | HP2-1 | 梅屋観光協会 | | HP2-2 | 梅屋周辺ガイド | 1 | | | |

【図８】

検索処理の処理内容を示すフローチャート



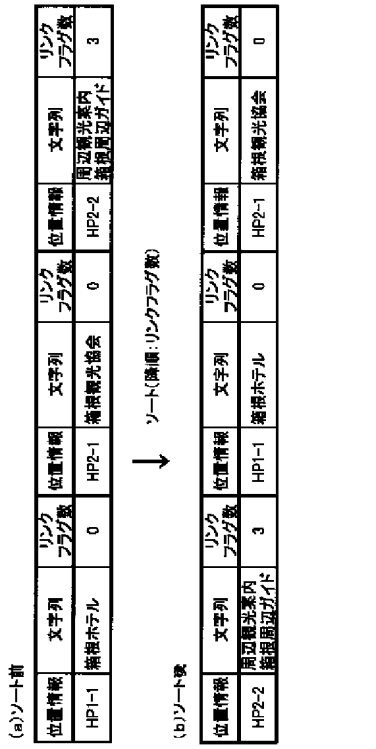

【図９】

検索結果ファイルの例を示す図

| 見出し語 | 位置情報 | 文字列 | リンクフラグ | 位置情報 | 文字列 | リンクフラグ | 位置情報 | 文字列 | リンクフラグ |
|---|---|---|---|---|---|---|---|---|---|
| 梅屋 | HP1-1 | 梅屋ホテル | | HP2-1 | 梅屋観光協会 | | HP2-2 | 梅屋周辺ガイド | 1 |
| ホテル | HP1-1 | 梅屋ホテル | | | | | HP2-1 | 梅屋観光協会 | |
| 観光 | HP1-1 | 梅屋ホテル | | HP2-2 | 周辺観光案内 | 1 | HP2-1 | 梅屋観光協会 | |
| 案内 | HP1-1 | 梅屋ホテル | | HP2-2 | 周辺観光案内 | 1 | | | |

【図１０】

検索結果リストファイルのソートの様子を示す図




【図１２】

記憶させた制御プログラムをコンピュータで読み取ることの可能な記録媒体の例を示す図